PUBLISHED FOR THE BROOKINGS INSTITUTION BY
JOHNS HOPKINS
UNIVERSITY PRESS

Brookings Papers ON ECONOMIC ACTIVITY

BPEA
FALL
2023

BROOKINGS

Janice Eberly and Jón Steinsson, Editors

Brookings Papers
ON ECONOMIC ACTIVITY

FALL 2023

Brookings Papers
ON ECONOMIC ACTIVITY

FALL 2023

JANICE EBERLY
JÓN STEINSSON
Editors

PUBLISHED FOR THE BROOKINGS INSTITUTION BY
JOHNS HOPKINS
UNIVERSITY PRESS

Brookings Papers on Economic Activity

1775 Massachusetts Avenue, N.W., Washington, D.C. 20036

ISSN 0007-2303
E-ISSN 1533-4465

Brookings Papers on Economic Activity is published twice a year in Spring and Fall by Johns Hopkins University Press, 2715 N. Charles Street, Baltimore, MD 21218-4363, USA. POSTMASTER: Send address changes to *Brookings Papers on Economic Activity*, Johns Hopkins University Press, Journals Division, P.O. Box 19966, Baltimore, MD 21211-0966.

The paper in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper). ♾

Brookings Papers
ON ECONOMIC ACTIVITY

FALL 2023

PURPOSE

Brookings Papers on Economic Activity (*BPEA*) publishes research on current issues in macroeconomics, broadly defined. The journal emphasizes rigorous analysis that has an empirical orientation, takes real-world institutions seriously, and is relevant to economic policy. Working drafts of the papers are presented and discussed at conferences held twice each year, and the final versions of the papers and comments along with summaries of the general discussions are published in the journal several months later. Research findings are described in a clear and accessible style to maximize their impact on economic understanding and economic policymaking; the intended audience includes analysts from universities, governments, and businesses. Topics covered by the journal include fiscal and monetary policy, consumption and saving behavior, business investment, housing, asset pricing, labor markets, wage and price setting, business cycles, long-run economic growth, the distribution of income and wealth, international capital flows and exchange rates, international trade and development, and the macroeconomic implications of health care costs, energy supply and demand, environmental issues, and the education system.

We would like to thank the supporters of the *BPEA* conference and journal, including the Alfred P. Sloan Foundation; General Motors Company; the National Science Foundation, under grant no. 2048708; Smith Richardson Foundation; and State Farm Mutual Automobile Insurance Company. We express appreciation to Brevan Howard Research Services for supporting *BPEA*'s mission and activities. We gratefully acknowledge Dina Axelrad Perry for establishing the George L. Perry and William C. Brainard *BPEA* Chair. We thank Janina Bröker, Aidan Kane, Georgia Nabors, Adam Sedlak, and Samuel Thorpe for preparing the summaries of general discussions for this volume.

 Conference drafts and recordings of the Fall 2023 *BPEA* Conference can be accessed at https://www.brookings.edu/events/bpea-fall-2023-conference/.

CALL FOR PAPERS

Most papers that are presented at the *BPEA* conferences and appear later in the journal are solicited by the editors, but the editors also consider unsolicited proposals. Editorial decisions are typically made nine months prior to each conference—proposals received by December 1 are considered for the following fall conference, and those received by June 1 for the spring. However, qualified proposals may be considered for conferences outside of the normal evaluation timeline, depending on the timeliness of the topics and the program needs. Proposals from early career researchers and members of underrepresented groups in the economics profession are encouraged. Proposals can be submitted at https://www.brookings.edu/bpea-for-authors/.

ACCESSING THE JOURNAL

All past editions of *BPEA*, including versions of the figures in color—along with appendix materials, data, and programs used to generate results—are made freely available for download at www.brookings.edu/bpea/search. To purchase print or e-subscriptions or single copies for institutions or individuals, please visit www.press.jhu.edu/journals or contact Johns Hopkins University Press, Journals Division, P.O. Box 19966, Baltimore, MD 21211-0966 (USA). Phone: 410-516-6987 • Toll free: 1-800-548-1784 (US/Canada only) • Email: jrnlcirc@jh.edu. Archived issues of *BPEA* are also available on Project MUSE (https://muse.jhu.edu/journal/52) and JSTOR (www.jstor.org).

ANNE CASE
Princeton University

ANGUS DEATON
Princeton University

Accounting for the Widening Mortality Gap between American Adults with and without a BA

ABSTRACT We examine mortality differences between American adults with and without a four-year college degree over the period 1992 to 2021. Mortality patterns, in aggregate and across groups, can provide evidence on how well society is functioning, information that goes beyond aggregate measures of material well-being. From 1992 to 2010, both educational groups saw falling mortality, but with greater improvements for the more educated; from 2010 to 2019, mortality continued to fall for those with a four-year degree while rising for those without; during the COVID-19 pandemic, mortality rose for both groups, but markedly more rapidly for the less educated. In consequence, the mortality gap between the two groups expanded in all three periods, leading to an 8.5-year difference in adult life expectancy by the end of 2021. There have been dramatic changes in patterns of mortality since 1992, but gaps rose consistently in each of thirteen broad classifications of cause of death. We document rising gaps in other measures relevant to well-being—background factors to the rising gap in mortality—including morbidity, social isolation, marriage, family income, and wealth.

Conflict of Interest Disclosure: This work was supported by a grant from the National Institute on Aging (NIA), given through the National Bureau of Economic Research (NBER). Anne Case serves on the National Advisory Council on Aging for the NIA. Other than the aforementioned, the authors did not receive financial support from any firm or person for this paper, or from any firm or person with a financial or political interest in this paper. Other than the aforementioned, the authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper. The discussant, Jonathan Skinner was previously a consultant for Sutter Health and is a program director at the NBER.

Brookings Papers on Economic Activity, Fall 2023: 1–44

Outcome gaps between adult Americans with and without a four-year college degree have become increasingly salient in politics, economics, demographics, and society more broadly. Voting patterns, wealth holdings, incarceration, wages, and marriage are now sharply different between the approximately one-third of the population age 25 and older with a bachelor's degree and the two-thirds without. Documenting differences in mortality between groups provides evidence on how well society is functioning beyond aggregate measures of material well-being. Compared with money-based measures of well-being, which depend on often controversial assumptions about what to include and on how to convert money into real measures, mortality is an objective measure, less subject to measurement error—someone is dead or alive—and there is little debate around which is better. Death is particularly indicative of societal failure when it is not due to a widespread infectious disease—like COVID-19—or even to failures in the medical system, but to self-inflicted causes like suicide, alcoholism, or drug overdose.

An examination of the mortality gaps between more- and less-educated Americans can tell us how the US economy is performing, not just on average, but for the majority of its population, those without a college degree. The division by education is in many ways an alternative to discussions of income distribution, for example by looking at outcomes at selected percentiles, and is a useful supplement to analysis by race and ethnicity. Educational differences are at least as salient as income differences. Similar considerations apply to international comparisons, where there has been much recent commentary on the superior economic performance of the United States relative to Europe, but where comparisons based on mortality are very different.

As we shall see in the next section, an examination of mortality for Americans with and without a college degree helps us understand the much-researched issue of why, since the 1980s, American life expectancy has performed so much worse than the life expectancies of other rich countries. This has been the topic of two reports from the National Academy of Sciences on international comparisons, Crimmins, Preston, and Cohen (2011) and Woolf and Aron (2013), as well as another more recent report on high and rising mortality in midlife in the United States, Harris, Majmundar, and Becker (2021). None of these reports focused on the mortality divide between those with and without a college degree.

We begin with an examination of life expectancy at age 25, often referred to as "adult life expectancy," which is defined as the number of years someone can expect to live beyond their 25th birthday if mortality rates were to

remain at their current levels. It is denoted by e_{25}. It is of course understood that mortality rates will change, and indeed the measure varies over time as mortality rates vary. It is *not* a forecast; like other period measures, it is a convenient summary of age-specific mortality rates, a single number that conveniently aggregates the many age-specific rates. Age 25 is taken to be the age by which people either have a four-year degree (BA) or will never have one, though see below.[1] In the next section, we show data on e_{25} for the United States and twenty-two other rich countries and how the differentials between Americans with and without a college degree help interpret the difference between the United States and other rich countries.

For technical reasons, which we shall explain as we go, most of the paper works with two other measures, expected years of life between the 25th and 85th birthdays, denoted ${}_{60}e_{25}$ (where the "60" refers to the number of years after age 25) and age-adjusted mortality between the same two birthdays. These two other measures ignore mortality rates after age 85. When there is no risk of confusion, we shall refer to both e_{25} and ${}_{60}e_{25}$ as *adult* life expectancy.

For the college-educated group, both measures of life expectancy at age 25 grew continuously from 1992 up to 2019, while for those without a four-year degree, progress stalled and reversed after 2010 (Sasson 2016a, 2016b; Hayward and Farina 2023). The gap widened further in 2020 and 2021 during the pandemic. We provide a descriptive analysis of the factors contributing to the widening gap in ${}_{60}e_{25}$ and in age-adjusted mortality, focusing on causes of death, on age, and on gender, both prior to and during the pandemic. We identify the causes of death that make the largest contribution to these widening gaps, particularly "deaths of despair"—from drug overdose, alcoholic liver disease, and suicide—as well as deaths from cancer, cardiovascular disease, chronic lower respiratory diseases, and diabetes.

The differential mortality experiences of those with and without a college degree come not only from direct effects of education on individual health, for example through health behaviors or enhanced ability to deal with life, including the health care system, but also from broader social and economic forces in the communities where people work and live. Those forces change with the structure of production and with the epidemiological environment, so that, for example, educational gaps in a service economy may be different from those in a manufacturing economy and may be different during a pandemic than before and after it. Who does or does not

1. A four-year college degree may be a bachelor of arts, science, fine arts, or architecture, among others. We use "BA" as a shorthand for all four-year degrees.

complete a four-year degree is also likely to depend on health, a selection effect. A good analogy here is with the college wage premium, the percentage by which the wage for college-educated workers exceeds the wage of those without a four-year college degree. This premium, which rose from 41 percent in 1979 to 80 percent in 2019, depends not only on what a college education does to the skills and ability of each worker—the direct effect—but also on a range of indirect effects, including on how many people go to college, who they are and how they are selected, for example, on ability; on how the labor market rewards skills; on available jobs and the technology they use; on how easy it is for workers to move to places where their skills are in demand; and on how the cost of employer health insurance affects the demand for more- and less-skilled workers (Finkelstein and others 2023).[2]

Similar direct and indirect forces affect health. Among them are the increasingly difficult job situation for less-educated workers and the long-term negative impacts of a deteriorating labor market on their marriages and the communities in which they live. (The recent tight job market has improved matters for less-educated workers (Autor, Dube, and McGrew 2023) but, as has happened in the past, the benefits may not last.) There is also important recent literature on the negative effects on health of corporate-sponsored laws passed in Republican-controlled state legislatures—regarding minimum wages, right-to-work laws, pollution, guns, and tobacco taxes and controls—all of which are likely to differentially hurt working-class Americans.[3]

European countries that have long been more open to trade and trade-related disturbances have built comprehensive welfare systems that help not only with trade-related job losses but also with losses through automation (Rodrik 1998). While mortality rates and mortality trends for less- and more-educated people in other rich countries differ in both levels and trends, the United States appears to be the only wealthy country where life expectancies are trending in different directions, one up and one down (Mackenbach and others 2018; Case and Deaton 2021, 2022).

2. The rise in the premium from 41 percent in 1979 to 80 percent in 2019 is from the authors' calculations using the Current Population Survey Outgoing Rotation Groups, for men and women age 25–64, comparing median wages for those with less than a four-year college degree to those with a BA or more. The premium for some college, less than a four-year degree, relative to a high school degree changed little over this period (14 percent in 1979, 12 percent in 2019).

3. See Grumbach (2022) and Montez and others (2020), as well as Jonathan Skinner's comment on this paper.

We document how gaps in mortality and life expectancy increased from 1992 to 2021, especially rapidly from 2019 to 2021 during the COVID-19 pandemic. We distinguish three periods: from 1992 to 2010, when both those with a BA and those without saw falling mortality, but with greater improvements for the more-educated; from 2010 to 2019, when mortality was falling for those with a BA and rising for those without; and from 2019 to 2021, when mortality was rising for both groups, but much more rapidly for those without a BA. We document the contributions of different causes of death to the changing gaps—notably the contributions of deaths of despair and their components, drug overdose, alcoholic liver disease, and suicide, and those of cardiovascular disease, and of a range of cancers—and we offer a complete accounting over all the major classifications of causes of death using the International Statistical Classification of Diseases (ICD).

In the final section of the paper, we turn from death to life and document the levels and trends in a range of outcomes for the more- and less-educated adult populations. Our underlying supposition is that the widening mortality gaps have their roots in differential life experiences between the two groups. Over a range of well-being–relevant outcomes, people with a college degree have fared better than those without. We do not attempt to link specific life outcomes to mortality rates, so we are *accounting* for the mortality outcomes only in the general sense of documenting the rising gaps in life outcomes among which, somewhere, lie the causal factors driving mortality differences.

We note that the fraction of the population with a four-year degree has risen over time. As is often discussed in the literature, rising educational attainment can change the kinds of people who do or do not have a four-year degree, a selection that can increase or decrease the educational gap in mortality and other outcomes, even when other effects of education are unchanging. We examine the mortality gap changes between birth cohorts when the fraction with a degree did not change, and again where the fraction with a degree changed markedly between birth cohorts. We find each successive birth cohort has a higher mortality gap than the cohorts that came before it, regardless of the change (or lack of change) in the fraction obtaining a degree.

We also show that reported educational attainment increases within birth cohorts, even long after the normal age of educational completion. Some of the increase can be accounted for by differential mortality but only for the earlier-born cohorts seen at older ages. The increase among the other groups remains a puzzle, and we can do no more than suggest explanations

such as adult education, immigration, or people as they age becoming more likely to claim having a degree when they do not.

There is a large body of literature examining the relationship between education and mortality, starting with Kitagawa and Hauser (1973). Many later studies focused, as we do, on changes in educational gaps over time; on identifying the causes of death underlying the gaps; on the differences between men and women, and between racial and ethnic groups.[4] Most recently, the perspective by Hayward and Farina (2023) emphasizes the contingent and changing nature of the relationship between education and mortality. From the earliest studies, cardiovascular disease and lung cancer were identified as important in explaining educational gaps, leading back to smoking as a key behavioral determinant, which itself differed for men and women both in prevalence and timing.

Educational attainment began to be recorded on the standard US death certificate in 1989, after which time, in principle, all decedents could be included in studies of education and mortality. Compared with mortality follow-ups using survey data, which have generated several important studies including Hummer and Lariscy (2011), Montez and others (2011), Montez and Zajacova (2013a, 2013b), the complete data permit the analysis of relatively rare causes of death, as well as disaggregation over a range of correlates. We use the death certificate information in this paper, and our work most closely follows earlier studies of the gap by Olshansky and others (2012), Meara, Richards, and Cutler (2008), and most recently and most closely, Geronimus and others (2019). Recent studies have documented that, particularly since 2010, drug overdoses, or more broadly deaths of despair, have become important in understanding the mortality gaps between those with and without a college degree (Case and Deaton 2017; Ho 2017; Sasson and Hayward 2019).

In the current paper, we update these studies in several ways and add a section on differential life outcomes other than mortality. We analyze annual data over the longer period now available, including the pandemic years 2020 and 2021. We choose a different, more limited, but sharper focus on the difference in outcomes between those with and without a four-year college degree. We are less concerned with the many possible mechanisms that account for the relation between education on health, and more with documenting differences in mortality associated with the college divide. This follows the analysis in our book *Deaths of Despair and the Future of*

4. See, for example, Preston and Taubman (1994) for an excellent early review and the more recent updates by O'Rand and Lynch (2018).

Capitalism (Case and Deaton 2020), where, among other things, we document the college divide in material well-being, morbidity, marriage, and religiosity. In the last section of this paper, we update these estimates for marriage and for morbidity, including mental distress, as well as for family income and wealth.

We use data for the thirty-year period from 1992 to 2021, though we go further back for some of the life measures whose deterioration traces back to the 1970s. The post-1992 period saw major changes in mortality patterns, including those for cardiovascular disease mortality—whose longstanding decline came to a halt—and those for several cancers, where there have been many improvements. Mortality from deaths of despair grew markedly over this period. We attempt to resolve some of the uncertainty about the relative contributions to declining life expectancy of changes in mortality from cardiovascular disease on the one hand and, on the other hand, rising mortality from deaths of despair, especially drug overdoses (Geronimus and others 2019; Mehta, Abrams, and Myrskylä 2020). The COVID-19 pandemic at the end of the period was characterized not only by COVID-19 deaths, but also by excess deaths from other causes, including an additional upsurge in deaths of despair. We document what happened to the mortality gap as mortality changed in these unprecedented ways.

We also use the classification in ICD-10 to offer a complete accounting of the contribution of all causes of death to changes in the gap and examine whether any causes of death act to reduce the mortality gap between those with and without a college degree. We ask if it matters for the gap whether the cause of death is one associated with rising mortality, falling mortality, or a change from falling to rising mortality. We also raise new questions about the measurement of educational attainment, adding to an ongoing debate about self-reports versus postmortem reports, a debate that has influenced the choice of data for studying the relationship between education and mortality.

I. Mortality: Data and Methods

In our analysis of mortality, we work with death certificates from 1992 through 2021, though in some cases we limit analysis to 1999–2021 so as to confine cause of death to the reporting structure of ICD-10, formally the International Statistical Classification of Diseases and Related Health Problems. Death certificates record age and sex, as well as highest education attained. We do not consider race or ethnicity in this paper but see Case and Deaton (2021), which documents the increasing importance for mortality

of education relative to race and ethnicity. There is undoubtedly some misreporting of education on death certificates, but the divide between a four-year college degree and less than a four-year college degree appears to be minimally affected (Rostron, Boies, and Arias 2010). As we shall document, there are also problems with self-reports of educational attainment. Education on death certificates is missing for four states in 1992: Oklahoma began reporting education in 1997, South Dakota in 2004, Georgia in 2010, and Rhode Island in 2016. These states accounted for 4.55 percent of the US population in 1990, and 4.57 percent of adult deaths in 1992. For deaths without education information, we assign a BA or not in the same proportion as nonmissing by year, age, and sex. Population totals for each year, age, and sex from age 25 to 84 are taken from the Census Bureau; the totals are split between those with and without a four-year college degree using ratios estimated from Current Population Surveys until 2000 and from the American Community Surveys thereafter. Our calculated statistics, age-adjusted mortality and adult life expectancy, are averages and as such reduce the influence of measurement errors.

We make extensive use of cause of death information as listed on the death certificates; we use the underlying cause of death, not proximate causes. The National Center for Health Statistics (2022, I.B, par. 2) notes "the underlying cause of death is the disease or injury which initiated the train of morbid events leading directly or indirectly to death or the circumstances of the accident or violence which produced the fatal injury." There is clearly scope for discretion and for error here, and causes of death are never as precise as the fact of death itself. There were particular difficulties during the pandemic, especially in the early days when testing was limited and when people died of other conditions that might not have proved fatal in the absence of COVID-19.

We use standard life table methods to calculate life expectancy at age 25, an age by which most people have completed their education; increasing attainment with age beyond 25 is an issue to which we return. The use of death certificates to compute mortality at the oldest ages is prone to error, and the official estimates from the National Center for Health Statistics (NCHS) use other sources (Arias and others 2022). We can avoid this by calculating the number of years of expected life of a 25-year-old between that person's 25th and 85th birthday, in standard demographic notation ${}_{60}e_{25}$, sometimes referred to as "temporary life expectancy" (Arriaga 1984). The standard measure of adult life expectancy e_{25} replaces 60 by infinity or at least the maximum possible years. Our measure of life expectancy from age 25 to 85 is also used by Geronimus and others (2019) who compute

expected numbers of years lost as 60 (the maximum possible number of years of life between age 25 and 85) minus expected life years, $60 - {}_{60}e_{25}$.[5]

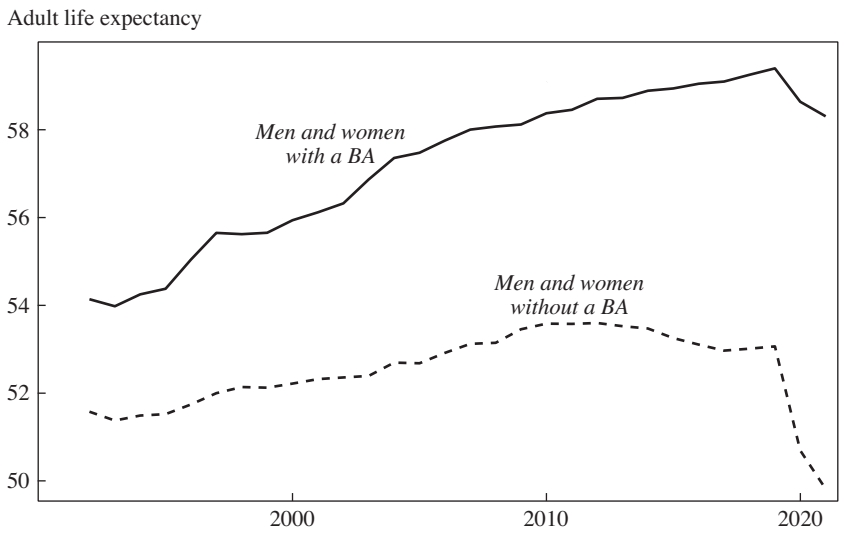
In the next section, we also report calculations of e_{25}. Here, too, we use the death certificates, extrapolating beyond age 85 using standard formulas that link mortality with age. We can provide some check on our calculations by using the same extrapolations to calculate e_{25}, not for those with and without a college degree, for which there are no official data, but for the whole population, and check against the official life tables, which we take in convenient form from the United States Mortality DataBase.[6] Our calculations are close to the official estimates; our maximum absolute error is 0.44 percent for women in 1992, and errors are smaller than that in later years, with maxima after 2000 of 0.27 percent for men in 2010 and 0.26 percent for women in 2021.

In section III and beyond, we make more complex calculations using individual causes of death, and we think it unwise to use interpolations to calculate those mortality rates at advanced ages; see above for the risk of errors at high ages. For these calculations we thus confine our attention to ${}_{60}e_{25}$ and to age-adjusted mortality between age 25 and 84. We compute age-adjusted mortality rates from age 25 to 84 for selected causes of death using the 2000 population and adjusting separately for men and women. We do not use separate reference populations by BA status; this is important because college graduates are on average younger than non-graduates, and we do not want these age differences to contribute to the gradient. We can use age-adjusted mortality rates, which are linear in both age-specific populations and causes of death, to exactly decompose the educational gaps by cause of death and by age group. For adult life expectancy, we use a variant of the cause deletion method (Beltrán-Sánchez, Preston, and Canudas-Romo 2008), in which we hold the age-, sex-, and education-specific mortality rates for selected causes at their 1992 levels, and then recompute adult life expectancy using the modified all-cause mortality rates. For example, deaths of despair rose rapidly after 1992, so to calculate the counterfactual excluding the increase, we compute ${}_{60}e_{25}$ as if that increase had not taken place, with all other mortality rates at their actual values. This is an accounting exercise, not a prediction of what would have happened. As was the case during COVID-19, deaths from other causes

5. We are grateful to John Bound for confirming that our calculations and those in Geronimus and others (2019) use the same formulas, something that is not clear in their text.

6. United States Mortality DataBase, University of California, Berkeley; usa.mortality.org (data downloaded on August 31, 2023).

Figure 1. Adult Life Expectancy for Americans with and without a Four-Year College Degree



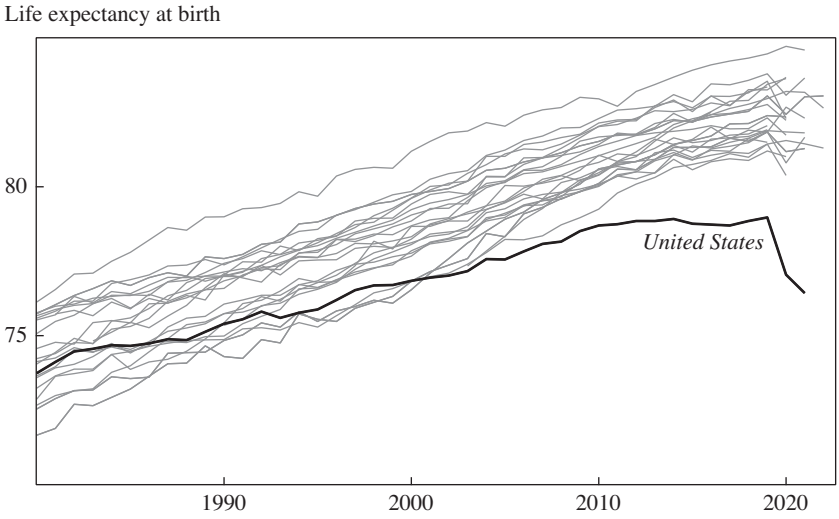
Source: National Vital Statistics System; and authors' calculations.

would almost certainly have been different had the increase in deaths of despair not happened; this is the well-known problem of competing risks, which precludes any straightforward, model-free calculation of counterfactuals. Even so, the calculations are useful in indicating orders of magnitude for the immediate consequences of modifying or eliminating different causes of death.

II. Adult Life Expectancy in the United States and Other Wealthy Countries

Figure 1 shows adult life expectancy, e_{25}, for Americans with and without a four-year college degree from 1992 to 2021; the figure combines men and women. The college-educated group experienced rising adult life expectancy until the onset of the pandemic in 2020. Those without a college degree saw their highest adult life expectancy in 2010 and have not regained it. Both groups lost years of life during the pandemic, 1.1 years for the college-educated, and 3.3 years for those without the degree. The gap widened throughout, from 2.6 years in 1992 to 6.3 years in 2019, and

Figure 2. Life Expectancy at Birth for the United States and Twenty-Two Other Rich Countries



Source: Human Mortality Database.

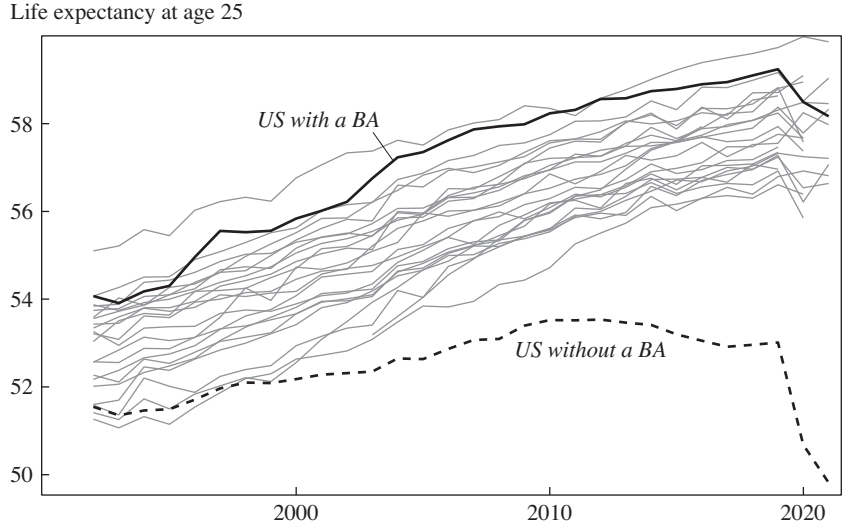
Note: The other countries shown in this figure, in order of their life expectancy in 2019, are Japan, Switzerland, Spain, South Korea, Italy, Australia, Sweden, Norway, France, Ireland, Canada, the Netherlands, Austria, Finland, Portugal, Belgium, New Zealand, Greece, Denmark, the United Kingdom, and Germany. The Israeli data end in 2016.

to 8.5 years in 2021 during the pandemic. (Note that at the time of writing, we cannot carry these calculations beyond 2021.)

We look at these results in more detail below, but we start by linking our findings to international comparisons between the United States and twenty-two other rich countries. Figure 2 shows a typical picture, here of life expectancy at birth, for the United States and for twenty-two other rich countries, with data taken from the Human Mortality Database.[7] In the mid-1980s, the US life expectancy at birth was in the middle of the range, but it has not kept up over time, and by the early 2000s, it was by far the lowest in the group. The pandemic added to an already large gap. The other countries shown in figure 2, in order of their life expectancy in 2019, are Japan, Switzerland, Spain, South Korea, Italy, Australia, Sweden, Norway, France, Ireland, Canada, the Netherlands, Austria, Finland, Portugal, Belgium,

7. Human Mortality Database, Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France); www.mortality.org (data downloaded on May 30, 2023).

Figure 3. Adult Life Expectancy for Americans by College Degree and for Twenty-Two Other Rich Countries



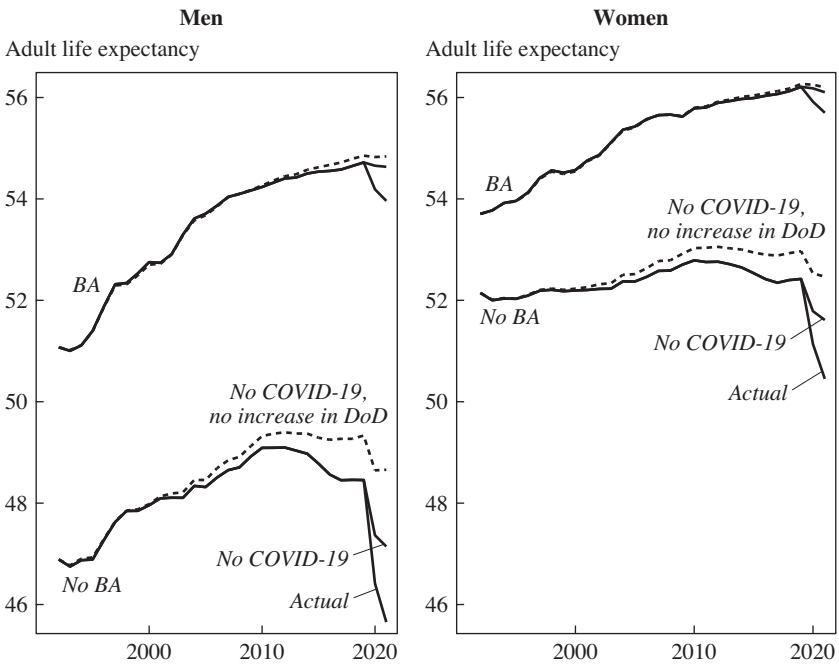
Source: Human Mortality Database; National Vital Statistics System; and authors' calculations.

Note: The other countries shown in this figure, in order of their adult life expectancy in 2019, are Japan, Switzerland, Spain, South Korea, Australia, Italy, Sweden, Norway, France, Canada, Ireland, New Zealand, the Netherlands, Austria, Finland, Belgium, Portugal, Greece, the United Kingdom, Denmark, and Germany. The Israeli data end in 2016.

New Zealand, Greece, Denmark, the United Kingdom and Germany. (The Israeli data end in 2016 with a life expectancy of 82.5 years.)

The literature lists many factors that can help explain the poor performance of the United States, and it is not our purpose to add to those accounts. Instead, we point to figure 3, which takes the Human Mortality Database data for adult life expectancy, e_{25}, for the other countries and superimposes the data from figure 1 of e_{25} for Americans with and without a college degree; this can only be done post-1992. One remarkable finding here is that Americans with a college degree, if they were a separate country, would be one of the best performers just below Japan, though there was some decline in 2020 and 2021 during the pandemic. We do not have life expectancy estimates by educational attainment for the other countries, though we do know that higher-educated people do better everywhere. But the figure shows that, without the widening gap in the United States, which is the main topic of this paper, the United States would not have done as relatively badly as it did.

Figure 4. Adult Life Expectancy with and without COVID-19 and Deaths of Despair (DoD)



Source: National Vital Statistics System; and authors' calculations.

III. Accounting for Education-Mortality Gaps in the United States

Figure 4 plots adult life expectancy from 1992 through 2021 for men and women separately, split between those with and without a BA. As noted above, we now work from here on not with e_{25}, but with ${}_{60}e_{25}$, the expected years of life between the 25th and 85th birthdays. If everyone died on their 85th birthday, the two measures would be identical. More generally, e_{25} exceeds ${}_{60}e_{25}$ by the product of life expectancy at age 85, e_{85}, and the fraction of those alive at age 25 who survive to age 85, quantities that have both been increasing as mortality rates have fallen, but both of which decreased during the COVID-19 pandemic. Since 1992, the difference $e_{25}-{}_{60}e_{25}$ (for both genders taken together and irrespective of degree status) has been between 1.9 and 3.1 years, rising from 1.95 in 1992 to 3.06 years in 2019 as mortality among the elderly fell and fewer adults died, and then falling to 2.46 years in 2020 and 2.34 in 2021.

The lower of each pair of solid black lines in each half of the figure is the actual outcome. For men with a BA, adult life expectancy, ${}_{60}e_{25}$, rose by 3.6 years from 1992 until 2019, from 51.1 to 54.7 years, then fell from 2019 to 2020 by 0.53 years, and again from 2020 to 2021 by 0.23 years. For women with a BA, our measure of adult life expectancy, ${}_{60}e_{25}$, rose by more than 2.5 years from 1992 until 2019, from 53.7 to 56.2 years, then fell from 2019 to 2020 by 0.29 years, and again from 2020 to 2021 by 0.22 years. Educated women gained less than educated men up to 2019 but lost less in the first two years of the pandemic. For men without a BA, adult life expectancy grew from 1992 to 2010 by 2.2 years, more slowly than for more-educated men over the same period, then fell by 0.6 years from 2010 to 2017, held steady for two years, and then fell dramatically during the pandemic by 2.0 years from 2019 to 2020 and by another 0.8 years from 2020 to 2021. For women without a BA, adult life expectancy grew from 1992 to 2010 by only 0.6 years, fell by 0.4 years from 2010 to 2017, held steady for two years, then fell during the pandemic by 1.3 years from 2019 to 2020, and by a further 0.6 years from 2020 to 2021. Once again, women gained less before the pandemic but lost less during it.

For both education groups, increases in life expectancy have been slower for women than for men. This is particularly dramatic for women without a college degree, for whom adult life expectancy in figure 4 in 2019, before the pandemic, was only 0.4 years higher than in 1992. Until the pandemic, men without a college degree had done better, gaining 1.5 years from 1992 to 2019, with all the gain coming before 2010. The slower gains for women are found in all rich countries, not just the United States. The main driver of mortality declines since the 1970s has been falling mortality from cardiovascular disease (CVD), primarily driven by reductions in smoking and by the use of antihypertensives and statins. But CVD is less prevalent among women who therefore had less to gain by the reduction. This effect is magnified by the fact that, in the United States as in most other countries, women were slower than men to start smoking and slower to stop, and smoking affects mortality not only through cancer but also through CVD.

The gap in adult life expectancy between the two education groups, which was 2.6 years (4.2 for men, 1.6 for women) in 1992, almost doubled to 5.0 years (6.3 men, 3.8 women) in 2019, and then exploded during the pandemic to 6.4 years (7.8 men, 4.8 women) in 2020, and 6.9 years (8.3 men, 5.2 women) in 2021. Accounting for these rising gaps is our main interest here.

The higher of the pair of solid black lines in figure 4, which differ from one another only in 2020 and 2021, shows the effects of eliminating

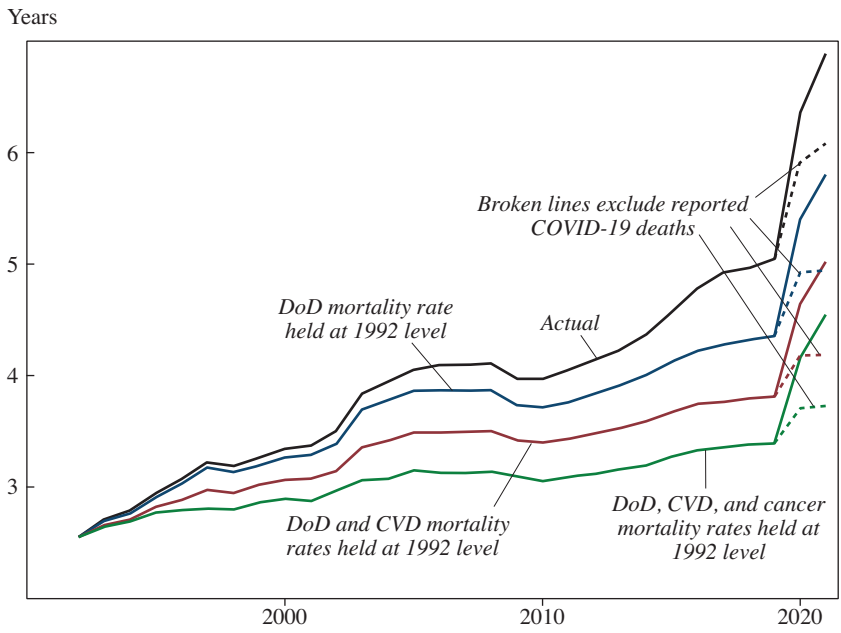
reported mortality from COVID-19; this deletion removes almost all of the drop for those with a BA, but only half the drop for those without. That excess deaths were greater than those reported as COVID-19 is well-known; the figure shows that the non-COVID-19 changes in mortality from 2019 to 2021, as well as the COVID-19 *excess* deaths in the pandemic years, were much larger for those without a BA. The higher dashed lines in both panels show estimates of adult life expectancy for each of the four groups when COVID-19 mortality is removed and the mortality rate from deaths of despair is held at its 1992 value. For those with a BA, the adjustment makes little difference beyond eliminating COVID-19 alone. For those without a BA, the actual and adjusted lines increasingly diverge as the epidemic of deaths of despair gathers momentum; indeed, the elimination of the increase in deaths of despair almost removes the post-2010 pre-pandemic decline in adult life expectancy for the less-educated group. It also moderates the declines during the pandemic; although the suicide rate fell in 2020, it rose again in 2021, and both drug overdose and alcohol-related liver disease mortality rates rose in both years.

Figure 4 also shows the three periods: up to 2010 when both groups were improving, but at different rates; from 2010 to the pandemic, when the groups were moving in different directions; and from 2019 when both groups were losing out, but at different rates.

Figure 5, for men and women combined, shows the evolution of the college gap from 1992 to 2021. The solid line marked "actual" is the gap; also shown are several counterfactuals. These include (1) eliminating COVID-19 deaths in 2020 and 2021; (2), as in (1) plus holding deaths of despair mortality rates at their 1992 levels; (3) as in (2) plus holding cardiovascular disease mortality rates constant at their 1992 values; then (4), all the above plus holding cancer mortality rates at their 1992 values. Each step reduces the temporal increase in the educational gradient. Note that both cardiovascular disease mortality and cancer mortality rates were falling over the period while the mortality rates from deaths of despair were rising. The figure does not show the effect on the *level* of life expectancy of, say, holding cancer mortality rates at their 1992 levels, something that would raise the mortality counterfactual in all subsequent years and lower life expectancy. Rather, the figure shows the effect of holding cancer mortality rates at their 1992 levels on *the educational gap* in life expectancy, and this, like the other counterfactuals, reduces the gap. In other words, the reduction in cancer mortality since 1992 has favored people with a college degree and has thus widened the gap.

Age-adjusted mortality data reproduce the qualitative patterns in figures 4 and 5 (online appendix figures 1 and 2). But because age-adjusted

Figure 5. Differences in Adult Life Expectancy with and without a BA



Source: National Vital Statistics System; and authors' calculations.
Note: DoD = deaths of despair; CVD = cardiovascular diseases.

mortality rates are linear in both age-specific mortality rates and population shares, they permit exact and straightforward decompositions by causes of death and by age groups. Table 1 presents pre- and post-pandemic age-adjusted mortality rates and covers eleven selected causes of death: deaths of despair, cancer, cardiovascular disease, chronic lower respiratory diseases, diabetes, transport accidents, Alzheimer's disease and related dementias, nephritis, septicemia, assault, and COVID-19. Collectively, these categories accounted for 80 percent of all deaths in 2019 for adults age 25 to 84. The ICD-9 and ICD-10 codes associated with these causes of death are listed in the notes to the table.

The first three columns of table 1 show age-adjusted mortality rates per 100,000 in 1992 for people age 25 to 84 with and without a BA, as well as the difference between them. The next three columns do the same for 2019, the last year before the pandemic. The next three columns show the changes from 1992 to 2019, so that the last column of this set shows the *differences in differences*, the changes from 1992 to 2019 in the gradient between those with and without a BA. The causes of death in the table are

Table 1. Age-Adjusted Mortality per 100,000 People, Age 25–84

	1992			*2019*			*Change 1992 to 2019*			*Change 2019 to 2021*		
Cause of death	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*
Deaths of despair[a]	26	43	17	29	95	66	3	52	49	3	37	33
Cancer	263	297	34	136	212	77	−127	−85	43	−5	−1	4
Cardiovascular diseases[b]	331	418	87	125	247	122	−206	−171	35	4	27	22
Respiratory[c]	33	50	17	16	55	39	−17	5	22	−2	−4	−2
Diabetes	18	28	10	13	33	20	−4	5	10	3	9	7
Transport	13	20	6	6	20	13	−7	0	7	0	5	5
Alzheimer's disease[d]	11	8	−3	23	28	5	12	19	7	1	4	2
Nephritis[e]	7	10	4	7	17	10	0	6	6	0	1	1
Septicemia	6	9	3	6	13	8	0	4	4	0	2	1
Assault	3	11	8	1	10	8	−2	−2	0	0	4	4
COVID-19	0	0	0	0	0	0	0	0	0	57	164	107
Total above[f]	710	895	184	362	730	368	−348	−165	184	63	247	184
Total mortality	845	1,056	211	462	908	445	−382	−149	234	66	265	198
Decomposition of deaths of despair												
Drugs	2	6	4	7	45	38	5	39	34	2	26	24
Suicide	13	16	3	11	22	11	−2	6	7	−1	1	1
Alcohol	11	21	10	11	28	17	0	7	8	2	10	8

Source: National Vital Statistics System; and authors' calculations.

a. Deaths of despair are from drugs (ICD-9 292, 304, 305.2–305.9, 850–858, 980, ICD-10 F11–F16, F18–F19, X40–X44, Y10–Y14); alcohol (ICD-9 291, 303, 305.0, 571.0–571.3, 571.5, ICD-10 F10, K70, K74.6, G31.2, X45, Y15); or suicide (ICD-9 950–959, ICD-10 X60–X84, Y87.0).

b. Cardiovascular diseases (ICD-9 390–459, ICD-10 I00–I99).

c. Chronic lower respiratory diseases (ICD-9 490–496, ICD-10 J40–J47).

d. Alzheimer's disease and related dementias (ICD-9 331.0, 290.0–290.4, ICD-10 F01, G30, G31.0, G31.1, G31.8, G31.9).

e. Nephritis, nephrotic syndrome, or nephrosis (ICD-9 580–589, ICD-10 N00–N07, N17–N19, N25–N27).

f. Differences in partial sums are due to rounding.

ordered by their sizes in this column. Finally, the last three columns present what happened during the pandemic, showing the contribution of each of the listed causes of death to the widening of the gap from 2019 to 2021.

In 1992, age-adjusted all-cause mortality rates for those with and without a BA were 845 and 1,056, respectively, a difference of 211. The corresponding figures for 2019 were 462 and 908, a difference of 445, an increase from the 1992 gradient of 234 age-adjusted deaths per 100,000. All-cause mortality fell between 1992 and 2019 for people with a BA, and more slowly from 1992 to 2010 for those without, rising thereafter. As a result, the gap in mortality between the two education groups increased from 1992 to 2019.

The eleven causes of death in table 1 account for 184 of these 234 deaths per 100,000, or 79 percent; a complete accounting for the period from 2000 to 2021 is provided below. The largest contribution comes from deaths of despair, which added 49 deaths to the change in the gradient, followed by cancer, 43, cardiovascular disease, 35, and chronic lower respiratory diseases, 22. The contributions of diabetes, transport accidents, Alzheimer's disease, nephritis, septicemia, and assault are smaller at 10, 7, 7, 6, 4, and 0, respectively. All estimates are rounded to whole numbers. This rounding accounts for any discrepancies in totals within the table. Apart from deaths of despair, where the increase in the gradient comes from a much larger increase in deaths among those without a college degree, the next largest increases in the gradient come from causes of death that have been falling over time.

The final three columns of table 1 track the changes in age-adjusted mortality rates and educational mortality gaps from 2019 to 2021. Three numbers are particularly notable. First, note the increase (from zero) of the number of deaths from COVID-19, and the very much larger age-adjusted mortality for those without a BA. COVID-19 alone added 107 age-adjusted deaths per 100,000 to the educational gap between 2019 and 2021. Second, there was a large increase in deaths of despair from 2019 to 2021, almost exclusively among those without a BA, 37 versus 3. Third, age-adjusted deaths from CVD also rose rapidly, again largely among those without a BA, 27 versus 4. Those three causes of death widened the gradient by 162, out of 184 for the causes of death shown in the table, and out of a total of 198 age-adjusted deaths from 2019 to 2021.

The last rows of table 1 decompose deaths of despair into its three components: deaths from drugs, from suicide, and from alcohol. All three have seen consistent increases in their contributions to the education mortality gradient since the early 1990s (see online appendix figure 3). Of the three, drug overdose is the largest contributor to the increase in the gradient and has received the most attention. But suicide and alcohol deaths have also

increased among those without a BA; particularly notable is the contribution of alcohol deaths to the increase in the gradient during the COVID-19 pandemic.

Table 2 shows a more complete characterization of causes of death from 2000 to 2021 using ICD-10 classifications; the shorter span of years obviates the need to match the classifications for ICD-9 and ICD-10. The table shows age-adjusted mortality rates for 2000 and 2019, as well as changes from 2000 to 2019 and from 2019 to 2021. Table 2 is constructed in parallel to table 1 but with different disease classifications. The text below the table explains the letter codes from ICD-10 and allows comparison of the two tables, despite the change in groupings. For example, deaths of despair in table 1 are now primarily captured in X and K codes. We have excluded causes that account for a small number of adult deaths so that columns 9 and 12 are now close to adding up to the totals in the last row, 137 out of 139 per 100,000 age-adjusted deaths for the change from 2000 to 2019, and 195 out of 198 per 100,000 for the pandemic years 2019 to 2021. Comparison of tables 1 and 2 shows that the former did not miss any diseases that made large contributions to the widening gradient, though table 2 identifies F codes (mental and behavioral disorders, some related to substance use), N codes (diseases of genitourinary system), A codes (certain infectious and parasitical diseases), and W codes (certain external causes, including falls) as making minor contributions to the widening gradients both before and during the pandemic.

An important result in table 2 is that, between 2000 and 2019, *all* causes of death, grouped by ICD-10 classification, contributed positively to the increase in the gap, and between 2019 and 2021, all except one did so, the exception being J codes, which cover deaths from respiratory diseases. This it true whether the mortality rate for the cause is falling for both groups (cancer, cardiovascular disease), rising for both groups (deaths of despair, respiratory diseases, Alzheimer's disease), or falling for the better-educated group and rising for the less-educated group (alcoholic liver disease, diabetes). With the one exception noted, the widening gap characterizes all time periods and all causes of death.

Figure 6 shows time series of age-adjusted mortality rates for age 25 to 84 for the three causes that contribute most to the increase of the gradient: deaths of despair, cancer, and CVD, by gender and by college degree status. Panels A and B show CVD mortality and deaths of despair, and panels C and D show cancer mortality. Panels A and B show that the rise in deaths of despair is more important for men than for women, and in both cases is almost entirely confined to those without a college degree. CVD mortality

Table 2. Age-Adjusted Mortality per 100,000 People, Age 25–84, by ICD-10 Category

	2000			*2019*			*Change 2000 to 2019*			*Change 2019 to 2021*		
Cause of death	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*	*BA*	*No BA*	*Diff*
COVID-19	—	—	—	—	—	—	—	—	—	57	164	107
X	16	34	18	21	77	56	5	43	38	1	31	29
Cancer	223	278	55	136	212	77	−88	−66	22	−5	−1	4
J	58	95	37	33	88	56	−25	−6	19	−3	−4	−1
Cardiovascular diseases	247	358	111	125	247	122	−122	−111	12	4	27	22
G	31	30	−2	44	49	8	10	19	10	2	5	3
K	23	40	17	19	44	25	−4	4	8	2	11	9
E	25	44	19	22	50	28	−2	6	8	3	12	9
F	9	13	5	15	26	12	6	13	7	1	4	3
V	11	20	9	6	19	13	−5	−1	4	0	5	5
N	12	21	9	10	23	13	−2	2	4	1	3	2
A	9	15	6	8	17	9	−1	2	3	1	2	1
W	7	10	2	9	14	5	2	4	3	1	2	1
Total above	672	957	285	444	867	423	−228	−90	138	65	260	195
Total mortality	702	1,008	307	462	908	445	−239	−101	139	66	265	198

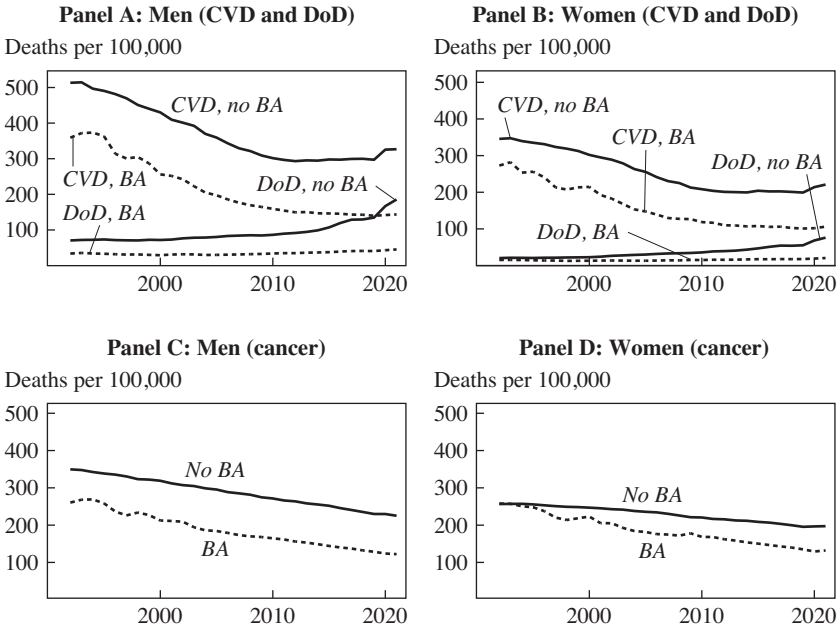
Source: National Vital Statistics System; and authors' calculations.

Note: ICD-10 codes:

X: Certain external causes, including accidental drug overdose, suicide, and assault with firearms
J: Diseases of the respiratory system, including chronic lower respiratory diseases, and influenza
G: Diseases of the nervous system, including Alzheimer's and Parkinson's diseases
K: Diseases of the digestive system, including alcoholic liver disease and cirrhosis
E: Endocrine, nutritional, and metabolic diseases, including diabetes
F: Mental and behavioral disorders, including those due to psychoactive substance use
V: Transport accidents
N: Diseases of genitourinary system
A: Certain infectious and parasitic diseases
W: Certain external causes, including falls

Certain causes of death are not shown in the table. These include ICD-10 codes B: Certain viral infections; D: Diseases of the blood and blood-forming organs; H: Diseases of the eye and adnexa, and diseases of the ear and mastoid process; L: Diseases of the skin and subcutaneous tissue; M: Diseases of the musculoskeletal system; O: Pregnancy, childbirth, and puerperium; P: Certain conditions from perinatal period; Q: Congenital malformations, deformations, and chromosomal abnormalities; R: symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere covered; and Y: Certain assaults, events of undetermined intent, sequelae of external causes.

Figure 6. Age-Adjusted Mortality Rates, Age 25-84, by BA Status

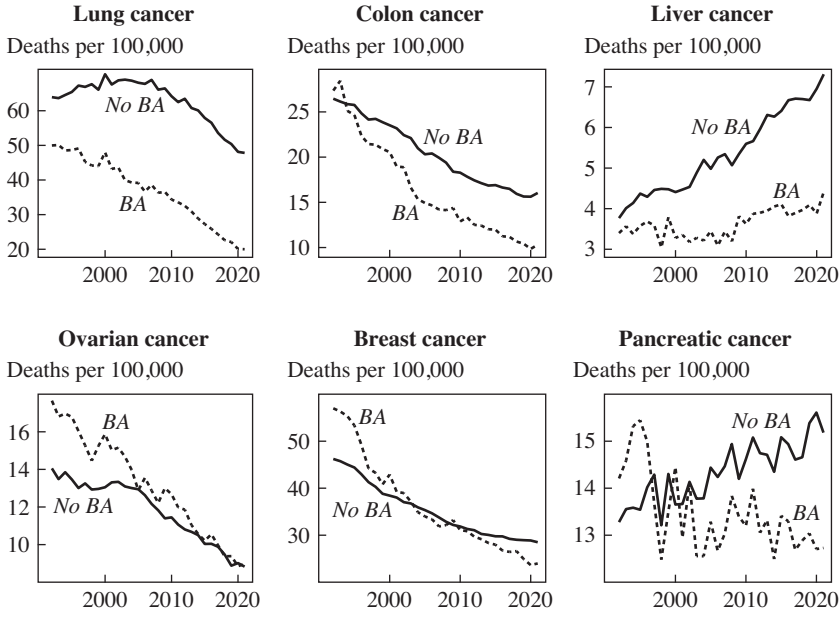


Source: National Vital Statistics System; and authors' calculations.
Note: DoD = deaths of despair; CVD = cardiovascular diseases.

also contributes to the widening gap for both men and women. The long-term decline that began in the 1970s lost momentum among those with a BA and stopped falling altogether after 2010 for those without the degree. After 2010, it rose slowly up to the pandemic and then more rapidly during it. These changes in the pattern of declining CVD mortality are recent, not well understood, and are of major importance not only for understanding the gaps but for understanding prospects for mortality more generally. Cancer mortality rates fell much more rapidly for women with a college degree than for women without. Indeed, in 1992, mortality rates from cancer were *higher* for more-educated women. For men, there is a more modest widening, with substantial decline for both those with and without a degree.

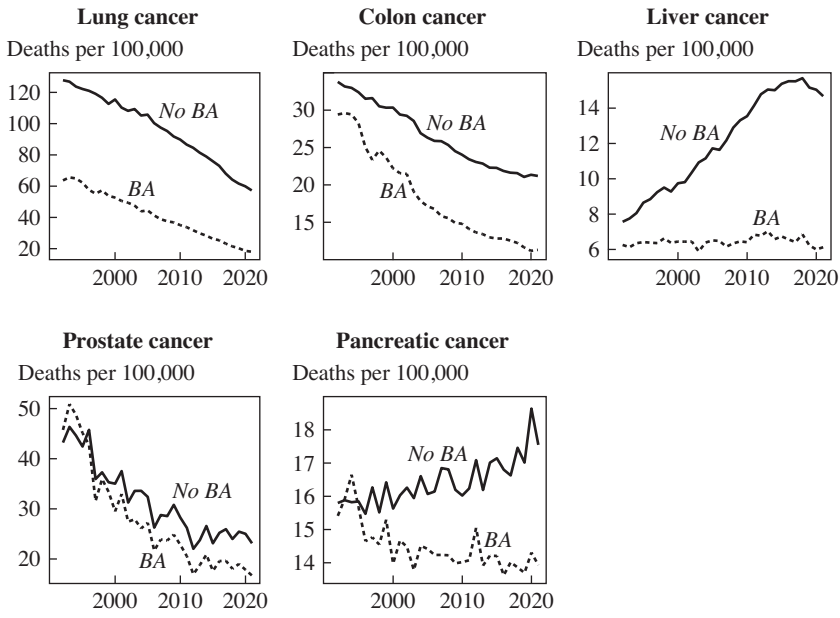
Figures 7 and 8, for women and men respectively, document patterns of mortality by education for the major cancers: for women, lung, breast, colon, ovarian, liver, and pancreatic cancer; and for men, lung, prostate, colon, liver, and pancreatic cancer. In the years immediately after 1992, lung cancer mortality was still rising for women without a BA but falling for those with

Figure 7. Age-Adjusted Cancer Mortality Rates, Women, by BA Status



Source: National Vital Statistics System; and authors' calculations.

Figure 8. Age-Adjusted Cancer Mortality Rates, Men, by BA Status



Source: National Vital Statistics System; and authors' calculations.

a BA. After 2006, lung cancer mortality fell for both groups in parallel, and since 2014 the gap has modestly narrowed. The contribution of lung cancer to widening the gradient for women comes before 2006. For men, who stopped smoking earlier than women, lung cancer mortality fell for both groups from 1992 to 2021, though more rapidly for those without a college degree, so that changes in lung cancer mortality for men worked to narrow the mortality gap. In 1992, breast cancer mortality was higher for women with a college degree, a long-standing finding that is often attributed to the protective effects of early childbearing.

But, as predicted by Link and others (1998), as scanning and effective treatment became available, breast cancer mortality fell more rapidly for the more-educated group who were first to use the technologies, contributing to a widening of the gradient. Prostate cancer mortality has fallen for men with and without a college degree, but more rapidly for those with, adding a relatively small amount to the widening of the mortality gap.

Among women, mortality from both colon and ovarian cancer were higher among those with a college degree in 1992, but as was the case for breast cancer, mortality fell more rapidly among women with a BA, crossing over for colon cancer and converging for ovarian cancer. As with breast cancer, screening and treatment were almost certainly both important. Mortality from liver cancer, whose risk factors include excessive alcohol use and cirrhosis, has been rising over time for both men and women, primarily among men and women without a college degree. Pancreatic cancer mortality has risen for both men and women without a college degree, while holding relatively steady after 2000 among those with a degree.

A key takeaway from figures 7 and 8 is that while different cancer mortality rates have behaved differently, with some falling and some rising, and while for some cancers mortality is or was higher for those with a college degree, for all the cancers examined here, with the exception of lung cancer for men, the educational gaps in mortality widened over time. Advances in medical treatments for many cancers and protective behavioral changes have had larger effects for those with a BA.

Table 3 calculates the college mortality gap by age group for 1992, 2019, and 2021. Column 1 gives the shares of each group in the population in 2000; these are the weights that can be applied to columns 2 through 6 to give the population totals in the bottom row. Column 2 gives the age-adjusted mortality rates in 2000 irrespective of educational status, while columns 3, 4, and 5 give the gaps—the differences in age-adjusted mortality rates between those with and without a four-year college degree. Column 6 shows the change in the gaps from 1992 to 2021; these changes are,

Table 3. College Gaps in Age-Adjusted Mortality by Age Group

Age group	*(1) Population shares in 2000 (%)*	*(2) Age-adjusted mortality rate[a] in 2000*	*(3) Mortality gap 1992[b]*	*(4) Mortality gap 2019[b]*	*(5) Mortality gap 2021[b]*	*(6) = (5)−(3) Change 1992 to 2021*	*(7) = (6)/(2) Change as a percentage of 2000 rate*
25–34	22.3	102	96	149	231	135	133
35–44	25.3	199	126	203	324	198	100
45–54	21.3	422	213	334	502	289	69
55–64	13.7	986	409	649	882	472	48
65–84	17.3	3,706	327	1,157	1,629	1,301	35
All 25–84	100	939	211	445	643	432	46

Source: National Vital Statistics; US Census Bureau; and authors' calculations.

a. Deaths per 100,000 persons.

b. Difference in mortality rate for people without and with a four-year college degree.

unsurprisingly, larger in groups with higher baseline mortality. Column 7 shows the changes as a percentage of the baseline mortality rates in 2000. The baseline of 2000 was chosen to align with its use in age standardization.

The overall increase in the gradient from 1992 to 2021 is 432 deaths per 100,000, to which the largest contribution comes from those age 65 and over, $(0.173 \times 1301)/432 = 52$ percent. The largest share of this is due to education differences in COVID-19 mortality, though there are also substantial contributions from cancer and CVD. As a percentage of baseline mortality, younger age groups saw larger increases in education gradients over this period; for the age group 25 to 34, the increase in the gap exceeded baseline mortality. Two-thirds of the increase among the youngest group was from deaths of despair. As we move from young to old, COVID-19 mortality becomes more important in contributing to the gradient, as does, to a lesser extent, mortality from CVD and cancer; deaths of despair become progressively less important with age.

IV. The Effects on Health of Education and of Rising Education

Our main interest is in documenting the changing differences in mortality between those with and without a four-year college degree, breaking up the patterns by cause of death, by gender, and by age. Our focus is not on the *reasons* for the better health of college-degree holders, which may include some or all of the following: (a) schooling in and of itself brings better health, better health behaviors, and better skills at dealing with health care, though the causal effect of education on health will always depend on the epidemiological environment, general health knowledge, and the structure of the health care system, as in fundamental cause theory (Link and Phelan 1995); (b) those who go to college are different in health-related ways, for example, in their health in childhood or in health-favoring personal characteristics (health-related selection) (Case, Fertig, and Paxson 2005; Farrell and Fuchs 1982); and (c) social and economic treatment if those without a college degree credential face a more difficult economic and social environment, including, for example, greater risk of job loss and community destruction. We might include in (c) the formulation, common in much of the sociological and epidemiological literature, that the main driver of health is socioeconomic status as measured by *rank* in the distribution of education (Adler and others 1994; Marmot 2004). It is only under (a) that we can argue that increasing the fraction with a BA might directly improve mortality rates; under (b) or (c) there is no such supposition.

Changes in health care provision, such as the Affordable Care Act (ACA), may differentially affect those with and without a BA, for example by increasing access to care among the less-educated group. We are skeptical that health care has large effects on population mortality rates, though in the diseases that we identify, the ACA may have reduced the gaps in cancer care and preventive treatment for cardiovascular disease. Given its design and purpose, the ACA surely played no role in *widening* the gaps.

Dynamic health-related selection can come into play when the fraction with a BA changes. Between 1992 and 2021, the fraction of the adult population age 25–84 with a BA or more rose from 22 to 36 percent. The increase for women, 18 percentage points, was larger than that for men, 10 percentage points, and these increases might contribute to the rising gap.[8] If the new college attendees are healthier than those who remain in the noncollege group, then a rising proportion of the population going to college will leave a noncollege group that is increasingly negatively selected on health. The effects of rising attainment on the educational health gap are not clear a priori because dynamic health selection as described will increase mortality rates for both groups, as the healthier nongraduates leave the pool of nongraduates, making the nongraduate group less healthy, and join an initially healthier graduate group, also reducing health in that group. (Despite the reduction in health in both groups, average health is unchanged.) As to the gap, it is straightforward to construct examples each of which yields different results. For example, if health h is uniformly distributed between zero and one, and those with $h > \theta$ go to college, a fraction $(1 - \theta)$, the average health in the two groups is $\theta/2$ and $(1 - \theta)/2$, and the gap is always 1/2, which does not depend on θ. If h has a standard normal distribution, the gap between average health of the college and the noncollege groups rises as the fraction going to college increases until half the population is in college and decreases thereafter. If h is exponentially distributed, the gap always *decreases* as more people go to college. Finally, if a subgroup of the noncollege people has poor health, and the rest are as healthy as the college group, then selection of the healthy previously noncollege group into college will have no effect on the average health of the college group, but it will decrease the average health of those not in college, thus widening the gap. Additional work and empirical evidence would be required to

8. Authors' calculations using the Current Population Survey Annual Social and Economic Supplement 1992 and the American Community Survey 2021.

document which, if any, of these illustrative calculations are relevant; we discuss the existing empirical evidence below.

No matter the effects of selection on the gap, the age-adjusted mortality rates and life expectancy numbers are not themselves affected. Selection does not challenge the facts, only their interpretation and what to do about them, if anything; this is not a situation in which selection leads to a biased estimate. In the extreme case where dynamic selection accounts for all of the increase in the gap, it might be argued that the widening is an inevitable and innocuous by-product of a desirable trend, the increase in education. We do not take this position, as we argue below, but simply note it.

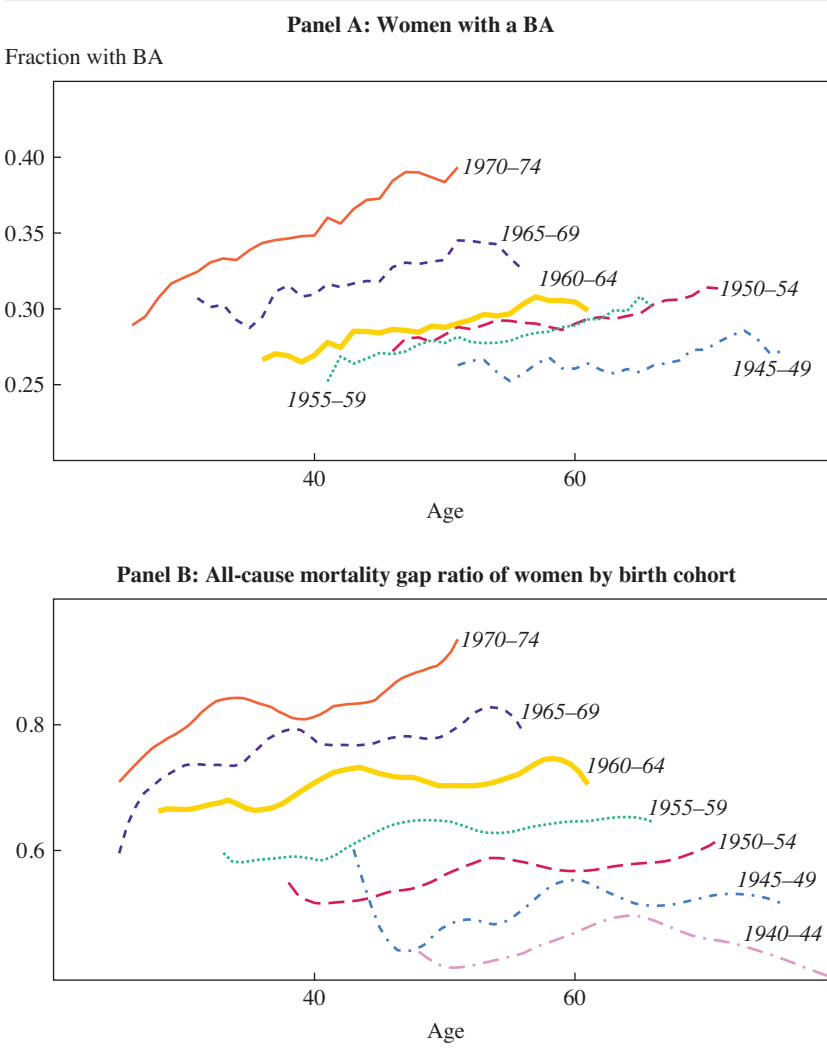
Several papers in the literature have made corrections for possible selection effects. Meara, Richards, and Cutler (2008) randomly reallocate some of their observations to keep constant the proportions in each of their groups. Others have worked with percentiles of the distribution of years of schooling, including Novosad, Rafkin, and Asher (2022) and Geronimus and others (2019) whose focus, similar to ours, is on mortality gaps between more- and less-educated Americans. While we look at people with and without a college degree, Geronimus and others (2019) compare outcomes for people in the bottom quartile of the education distribution with those in the top three quartiles. Even if educational qualifications were measured continuously, it is unclear why what happens at a particular percentile is of interest given that jobs and social standing depend more on qualifications than on percentiles, nor how, in the presence of health selection, looking at percentiles identifies a specific parameter of interest. Geronimus and others (2019) assign quartiles within (birth year, sex, race) cells for Black and white non-Hispanic individuals. For white non-Hispanic individuals, examination of the data shows that the bottom quartile has been defined by a high school degree since the birth cohorts born in the early 1920s, and for Black individuals since the early 1940s. As a result, a comparison of the bottom quartile to the rest of the distribution is similar to a comparison between those with no more than a high school degree to those with at least some college education. Their categorization differs from ours, in practice, in allocating the group with some college but less than a BA to their "high" education category. In previous work, we have shown that socioeconomic outcomes and mortality patterns for those with some college but no BA are closer to those with a high school degree or less than to those with a college degree. (We update and explore this in online appendix figure 4.) Despite this difference, their estimates are qualitatively similar to ours. That this is so provides evidence that the selection effects on the gap are not very important.

Similar and even clearer evidence comes from Novosad, Rafkin, and Asher (2022) who believe that it is educational rank that matters, not educational attainment, and who develop a method of estimating mortality change over time at fixed percentiles by age, race, and sex. Because educational attainment is discrete rather than continuous, it is only possible to estimate mortality change within an interval, but the payoff to the method is that the selection effects are eliminated by holding percentiles constant. Figure 6 in Novosad, Rafkin, and Asher (2022), for 1992 to 2018, shows very large percentage increases in mortality for white males and females below the 10th percentile (these are primarily high school dropouts). They also show mortality increases for those under age 50 that extend in some cases up to the 70th percentile, essentially to everyone without a BA. Setting aside the broad issue of whether it is qualifications or ranks that matter, these estimates eliminate health-based selection into education and so provide direct evidence that (fixed groups of) less-educated Americans have seen substantial mortality increases while those with the highest education levels have seen a continuing mortality decline.

Yet more evidence comes from examining the changes in college completion and changes in mortality gaps for women in the United States born between 1940 and 1974, using the fact that women's college completion did not increase by the same amount between successive birth cohorts. Panel A of figure 9 presents the fraction of women who completed a BA in each of seven five-year birth cohorts from 1940–1944 through 1970–1974, using data drawn from the American Community Survey. College completion increased between the cohort of 1940–1944, when approximately 21 percent of women completed a BA, to the cohorts of 1945–1949 (26 percent) and 1950–1954 (28 percent). There was then a period of stagnation for the birth cohorts from 1950–1954 through 1955–1959 and 1960–1964, after which the upward trend in the fraction of women with a BA in successive birth cohorts resumed. Explanations for rising mortality gaps that rely on selection would suggest that increases in mortality gaps between the first and second birth cohorts (1940–1944 and 1945–1949) and the increases between the last three cohorts (between 1960–1964 and 1965–1969 and between 1965–1969 and 1970–1974) should be larger than those for women born at midcentury.

We look at this using a relative mortality gap measure. For each age and year, we calculate the mortality gap ratio $(m_{noBA} - m_{BA})/m_{ALL}$, the mortality difference between those without a BA and those with a BA or more, scaled by the mortality rate for the population of the whole cohort. Note that this measure is corrected for any age effects that affect numerator and

Figure 9. College Completion and Mortality Gap Ratios of Women by Birth Cohort



Source: American Community Survey 2000–2021 (panel A); US National Vital Statistics 1992– 2021 (panel B).

denominator proportionately. We present these gap ratios for each of the seven birth cohorts in panel B of figure 9, where we smooth between ages within each birth cohort using a second-degree polynomial smoother. Contrary to what would be expected if selection were the driving force in mortality gaps, we find that the gap ratios rise by approximately 6 percentage points between each cohort; there is no pause for the cohorts born at midcentury when women's college completion rates did not change. In contrast, for later- and earlier-born cohorts, for whom education increased substantially, the increase in the mortality gap ratios was similar to that for the cohorts where education was not changing. In general, the upward movement in the mortality gaps appears to have no relation to changing fractions with a BA.

Men's college completion followed a different path between birth cohorts. Attainment of a BA rose between the birth cohort of 1940–1944 and that of 1945–1949. However, this was followed by a drop in the rate of college completion, a drop that held through the birth cohort of 1960–1964, after which the fraction of cohort members with a BA began to rise in successive birth cohorts. Once again, this pattern is not matched by that found in the mortality gap ratios for men, which follows the pattern observed for women in panel B of figure 9. Successive birth cohorts of men from those born in the early 1940s to those born in the early 1970s have seen increases in the mortality gap ratio that average 5.5 percentage points between birth cohorts, regardless of the fraction of the cohort with a BA (see online appendix figure 6).

Finally, we note that in their review of the literature on education and mortality, Hayward and Farina (2023, 401) conclude that "although selection cannot be completely ruled out, most of the evidence runs counter to what one would expect given negative selectivity." Our evidence supports that conclusion. We are unaware of any studies to the contrary that show dynamic health selection to be quantitively important.

Examination of educational attainment within each birth cohort shows that the fraction of those reporting a college degree increases as the cohort ages. For example, for those born in 1940, a regression of degree attainment on age attracts a coefficient of 0.0011, so that between when we first see them at age 52 and last see them at age 81, the fraction with a college degree has increased by more than 3 percentage points. For younger cohorts, the numbers are larger; for example, for the cohort born in 1970, the fraction reporting a degree increases by 14 percentage points from age 25 to 51. Differential mortality rates—which we have in our data—will differentially select out the less-educated as each cohort ages, but this effect

is negligible for the younger cohorts. For the cohort born in 1940, differential mortality should increase the fraction with a degree by 4 percentage points, but for the 1970 cohort, the increase is less than 1 percentage point. According to the National Center for Education Statistics, about a quarter of college graduates in 2012 obtained their degree between age 25 and 39, presumably mostly at the lower end of that range.[9] Even so, there is upward drift within cohorts beyond age 30 (and even beyond age 40) in the reported fraction of degree holders.

The upward drift in reported possession of a bachelor's degree for later-born cohorts cannot be explained by differential mortality and is unlikely to be fully explicable by people going to college at later ages. Immigrants are about as likely as native-born Americans to have a college degree (Krogstad and Radford 2018), and results on upward drift are similar when we restrict our sample to the native-born population, so we are left with the supposition that people are granting themselves degrees as they age. There are certainly great incentives to do so, and perhaps few risks to people checking a box on a website for jobs in the hope that prospective employers will not check.

What does this imply for the analysis in this paper, or indeed for other papers in the literature that assume that education is complete by age 25? Effects ascribed to having a college degree are, at least in part, confounded with the effects of compositional change, even within birth cohorts. Several papers have questioned the use of education as reported on death certificates on the grounds that it is *not* self-reported and have taken that as a reason to work with the (much smaller) mortality follow-up of the National Health Interview Survey (Hendi 2017; Masters, Hummer, and Powers 2012). Yet our results show that self-reports may also be problematic. If the main concern is adults going back to college, the analysis can be confined to those age 35 (or 45) and above, and we note that figures 4 and 5 show the same patterns of widening gradients if we work with ${}_{50}e_{35}$ or ${}_{40}e_{45}$ in place of ${}_{60}e_{25}$. Our parallel with calculations of the college wage premium is unaffected in the sense that the health and wage premia are both based on potentially exaggerated degree attainment. Each should be interpreted as the difference in earnings or mortality outcomes between those who have or claim to have a college degree and those who do not. Many people who

9. National Center for Education Statistics, "Integrated Postsecondary Education Data System," https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?year=2012&surveyNumber=3&gotoReportId=7&, accessed April 9, 2023.

falsely claim to have a degree may still receive at least some of the social and economic benefits of having one.

V. Mortality: Discussion

Meara, Richards, and Cutler (2008) examine mortality by education up to 2000 and entitle their paper "The Gap Gets Bigger." Their title works just as well for the mortality gap between Americans with and without a bachelor's degree in the subsequent years, from 2000 to 2021. Indeed, the rate of widening accelerated after 2010 and exploded during the pandemic.

The years between 1992 and 2021 were years in which patterns of mortality changed dramatically, and those changes were different for men and for women. What is remarkable is that the widening of the gap transcended these changes in the mortality patterns. This would have been remarkable enough for the gap in all-cause mortality as the underlying causes of death changed. What is more surprising is that the widening gap is seen in virtually all the major groupings of causes of death. We see it in deaths whose rates have risen in the last thirty years, like deaths of despair and COVID-19; we see it in deaths whose rates have fallen in the last thirty years, like cancer; we see it in deaths whose rates have fallen and then risen, like deaths from cardiovascular disease; and we see it in deaths whose rates were originally higher for those without a BA (most diseases) and those that were originally lower for those without a BA (colon, liver, ovarian, and breast cancer for women, and prostate and pancreatic cancer for men). Even though the mechanisms and stories are different for each disease, and sometimes different for men and women, the widening gap is almost always there.

The words *virtually* and *almost* are there to note the only exception that we found, which occurred during the two-year period from 2019 to 2021 for the category of ICD-10 labeled "diseases of the respiratory system, including chronic lower respiratory diseases, and influenza," which excludes deaths from COVID-19. From 2000 to 2019, the gap in this category widened, as in other causes of death. During 2020 and 2021, the pandemic years, some respiratory diseases may have been misclassified as COVID-19 and, given that COVID-19 deaths were much more common among those without a BA, the narrowing of the gap in respiratory diseases could be due to misattribution. Note again our earlier comments on the difficulties of assigning cause of death in such complex cases.

We note too that while an increasing mortality gap is seen in cancer as a group, the gap is shrinking for one specific cancer, lung cancer. Men

with a BA gave up smoking much earlier than men without, but in the past thirty years the latter have been quitting too, which has narrowed the gap for men. For women, the mortality gap in lung cancer increased until 2006 before stabilizing, while continuing to increase in other cancers.

Fundamental cause theory says that, whenever there exists the means to prevent death, those means will be more effectively seized by those with power and resources (Link and Phelan 1995). What we are seeing here are fundamental cause mechanisms on steroids; the gap is not just present but expanding, and expanding at an accelerating rate. Either the gap in power and resources is expanding or the means of preventing disease has been growing; we suspect both are true. We do not have a well-documented account of how and why this is happening, but point instead to the fact that these gaps between those with and without a BA are widening across a range of life outcomes that we have reason to care about, not just mortality, but also morbidity—including many kinds of pain—as well as marriage rates, childbearing outside of marriage, religious observance, institutional attachments, and wages and participation in employment.[10]

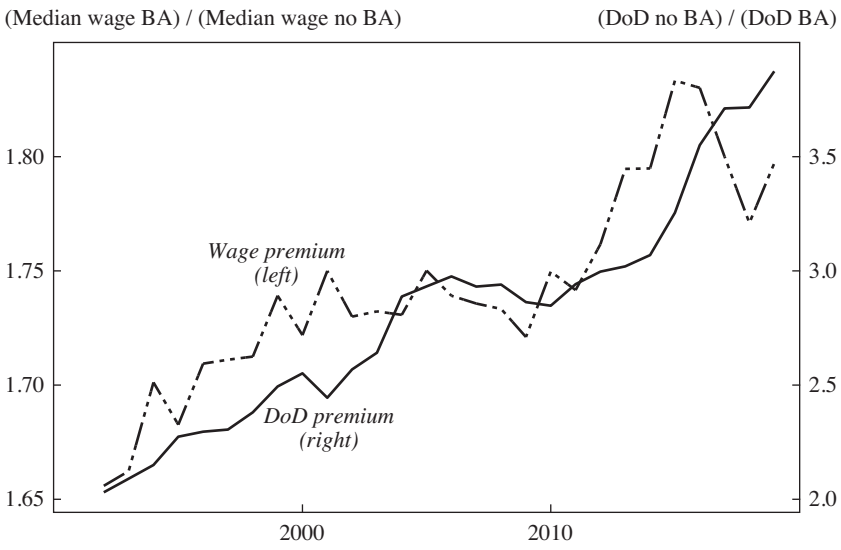
Figure 10 sets the stage for section VI and illustrates with one such comparison, between wage rates and deaths of despair. The dotted line (left-hand axis) shows the college wage premium defined as the ratio of median wages for those with a BA or more to median wages for those without a BA, while the solid line (right-hand axis) shows the ratio of the age-adjusted mortality rate from drugs, alcohol, and suicide for those without a BA to the age-adjusted mortality rate for those with a BA or more. In both cases, we look at age 25 to 64. Note that we are not arguing for a direct causal connection here; instead, we think of these series as two of many ways of documenting the deterioration in the situation of less-educated people in today's United States. Note that both comparisons show rising gaps up to 2000, then a period of relative pause, followed by an acceleration after 2010. A closing of mortality gaps may be an elusive goal while gaps in other domains continue to increase.

VI. Gaps among the Living

The decades-long increase in mortality gaps we have documented are matched by widening gaps in many measurable outcomes among the living, of which figure 10 is one example. We do not try to pin causality on any of the measures we document, though differences in adult mortality,

10. See Case and Deaton (2020) and section VII.

Figure 10. Ratios of Median Wages (BA/no BA) and Age-Adjusted Mortality Rates, Drugs, Alcohol and Suicide (no BA/BA), Age 25–64



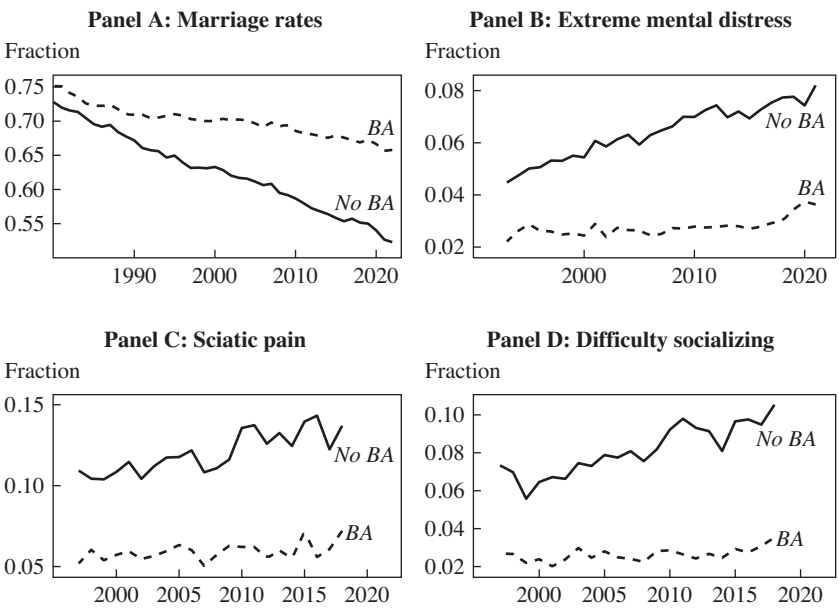
Source: CPS Outgoing Rotation Groups; National Vital Statistics Study; and authors' calculations.
Note: DoD = deaths of despair.

especially differences in mortality that are essentially self-inflicted, are certainly rooted in differences in the lives that preceded them. In such accounts, causality would certainly operate slowly and cumulatively or, to borrow a phrase, with long and variable lags. We do not attempt to disentangle the potential roles of the factors we consider in affecting either deaths of despair or overall mortality. That said, we note the excellent work on the precursors of deaths of despair by Olfson and others (2021). Merging individual data from the American Community Survey with death records, Olfson and others (2021) report the risk of dying from drugs, alcohol, or suicide (each analyzed separately) is higher for those who are single, those who have less than a four-year degree, and those who report lower income; they show that the difference between people with and without a BA remains after controlling for a number of other factors.

We examine gaps and changes in gaps by BA status in marriage, social isolation, pain, mental health, income, and wealth. Our findings parallel the earlier documentation of gaps in mortality in that the gaps between those with and without a BA have been widening since at least the mid-1990s.

Figure 11 plots marriage rates, as well as rates of physical pain and mental distress. All are age-adjusted to the 2000 US population and combine men

Figure 11. Marriage Rates and Rates of Physical and Mental Distress



Source: Current Population Survey 1980–2022 (panel A); Behavioral Risk Factor Surveillance System 1993–2021 (panel B); National Health Interview Survey 1997–2018 (panels C and D); and authors' calculations.

and women age 25 to 79. The pain measure relates to sciatic pain—a type of pain that is specific and likely reliably reported. It and the fraction of people who report that they have difficulty socializing ("visiting friends, attending clubs") come from the National Health Interview Survey (NHIS) and run from 1997 to 2018; the NHIS was redesigned after 2018, and the later data are not comparable. The "difficulty socializing" measure captures one aspect of loneliness, a condition recently described as an epidemic by the Surgeon General of the US Department of Health and Human Services (2023); the standard surveys on which we rely do not have the more sophisticated questions that would be preferable.

The measure of extreme mental distress comes from the Behavioral Risk Factor Surveillance System (BRFSS) and was first suggested and used by Blanchflower and Oswald (2020) to analyze educational differences in mental health. The question asks, "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?" The graph plots the fraction of the population who replied thirty days, that is,

whose mental health was not good on every day of the past thirty. Finally, marriage rates are taken from the Current Population Survey.

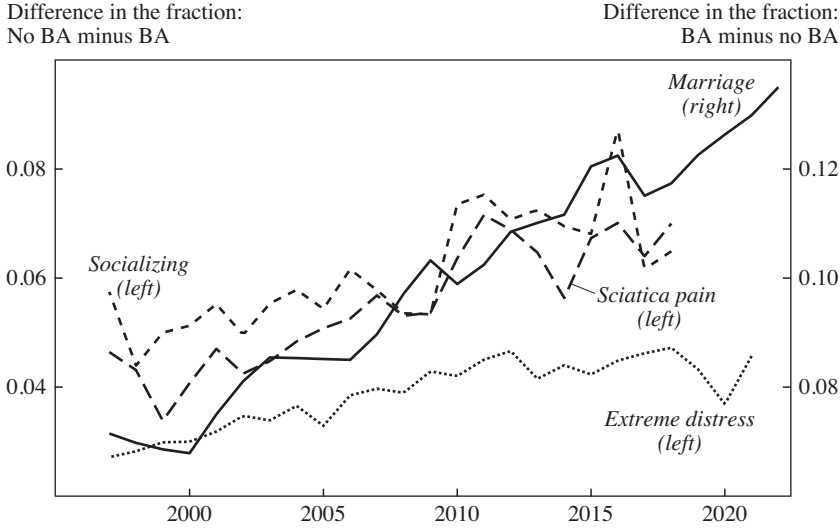
The fraction of adults currently married has been declining for those without and with a BA. From 1980 to 1990, the two lines fell in parallel, but since then, the fall has been markedly more rapid among those without a college degree (see online appendix figure 5). The decline persisted and perhaps slightly accelerated during and immediately after the COVID-19 pandemic. The long-established decline has been explored in the sociological literature on "fragile families," which describes the still-increasing phenomenon of serial cohabitation, often with children, who then live separated from one or the other of their parents (McLanahan 2004; Cherlin 2014); the decreased attachment to the institution of marriage is part of a wider detachment from social institutions, including religion, by working-class Americans (Edin and others 2019).

The other three measures in figure 11 are all rising over time, getting worse for both educational groups, but the increase is much more pronounced for those without a four-year college degree. Extreme mental distress has risen steadily since the early 1990s for those without a college degree and by little for those with a degree before 2015. In 2019 to 2020, and 2020 to 2021, the two groups moved in opposite directions, down and then up for the less educated and up then down for those with a BA. These contrary movements during the COVID-19 pandemic are worth further analysis. The measures of sciatic pain and of difficulty socializing come from the NHIS whose sample size is smaller, and are relatively noisy; even so the greater prevalence of both among the less-educated is clear. As reported in Lamba and Moffitt (2023), the largest increase in reported pain occurred for those without a BA during the financial crisis, and the increase in this gap persisted through 2018.

Figure 12 summarizes the gaps in single picture in which the gaps for all four measures are rising over time. This graph shows a parallel with our findings on mortality in that the gaps between the two groups have grown and are growing over time. Of course, we should not push the analogy too far; all four of the measures here are worsening over time, while several of the mortality rates, particularly for cancers, were improving.

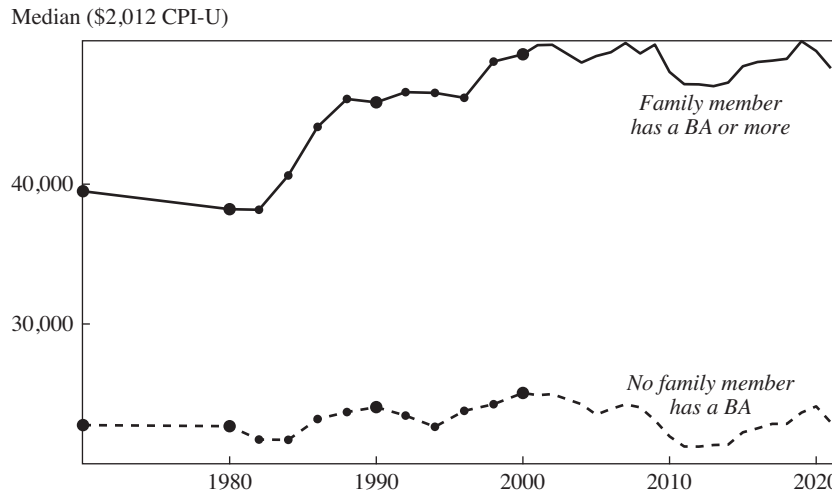
When we turn to income and wealth, the general trends are of improvement, albeit at different rates for the two groups. Figure 13 shows real family incomes from 1970 to 2021; 1970 is often identified as the year after which broadly shared general prosperity broke down. The data come from the US Census in 1970, 1980, 1990, and 2000, shown as large dots; from the CPS for the non-census years from 1980 to 1999, shown as smaller dots;

Figure 12. Education Gaps in Marriage and Mental and Physical Health



Sources: Authors' calculations using data in figure 11.

Figure 13. Real Family Total Income per Equivalence Unit ($2,012 CPI-U)



Source: US Census (1970, 1980, 1990, 2000); Current Population Survey (1982–1988, 1992–1998); American Community Survey (2001–2021); and authors' calculations.

and from the American Community Survey annually since 2001. We have deflated by the Consumer Price Index for All Urban Consumers (CPI-U) to real 2012 dollars and calculated family equivalents in which each child under 18 counts as 0.7 of an adult and where the sum of adults plus 0.7 children is raised to the power of 0.7 to capture economies of scale.[11] If we were to use the price deflator of per capita expenditure in place of the CPI-U, both income measures would rise somewhat more rapidly, though the change in the gap does not change qualitatively. There is scope for much argument about the choice of price indexes, but the main difference between the two is different weights, with the personal consumption expenditures (PCE) deflator including many items that families do not directly purchase.

The headline from this figure is that the gap in real equivalized family income increased, from $16,500 in 1970 to more than $25,000 in 2022. The increase was not steady over the half century shown. It fell slightly from 1970 to 1980, rose rapidly in the 1980s, rose more slowly from 1990 to 2010, and has been trendless since. We know the underlying anatomy of these changes. Part is the increase in the college wage premium, from 41 percent in 1979 to 80 percent in 2019.[12] The 1980s and, to a lesser extent, the 1990s were also periods of rising family income inequality, to which the gap between the education groups contributed. The changes also reflect rates of labor force participation that differ by educational status, as well as by men and women. For those without a BA, the employment-to-population ratio for men has been falling, albeit with cyclical interruptions, since 1980, while for women, the ratio rose until 2000 and fell thereafter. For men and women with a BA, the patterns are similar, but the increases and decreases are much smaller. As a result, differential participation rates contribute to widening the gap until around 2010. In the recovery from the pandemic, these patterns have changed, with better outcomes for low-skilled workers, but it is too early to tell whether the long-term pattern has changed. To the extent that the increase in employment by less-educated women after 1970 was a compensatory, but sometimes unwelcome, response to falling real wages by men, changes in family income may overstate changes in well-being.

We have not attempted to adjust the gaps for taxes paid—these are pretax incomes, though they include benefits such as unemployment compensation,

11. See Citro and Michael (1995) for this and other measures.

12. Authors' calculations of the wage premium, measured as the ratio of median real wages for those with a BA to median real wages for those without a degree, for workers age 25 to 64 in the Current Population Survey Outgoing Rotation Groups.

workers' compensation, supplemental security income, and public assistance or welfare payments. Nor do we adjust for any increase in quality that is missed in the CPI, let alone for possible differentials in the rates of quality improvement between groups. We do not include employer contributions to health insurance as income; we note that those are not very different for less- and more-educated workers, though there are presumably differences by employment. Given that those with a BA are more likely to have such coverage, incorporating such contributions would increase the gap. We do not attempt to put a value on coverage nor to subtract out the part of costs that is due to health care industry rents. Nor, finally, do we add in the value of Medicaid and Medicare as some have argued for (Burkhauser and others 2024). Corrections of this kind, if indeed they can be justified as corrections, would have uncertain effects on the gap, although they would undo some of the stagnation of real incomes among families without a BA.

Wealth data from the Survey of Consumer Finances can be used to study differences by education. In particular, the infographic provided by the Board of Governors of the Federal Reserve System showed (as of July) that, taking all components of household wealth together, the total in 1990:Q1 was $20.91 trillion, rising to $140.56 trillion by 2023:Q2.[13] In 1990, the fraction owned by those without a college degree was 49 percent, a fraction that had fallen to 27 percent by 2023, so that those with a college degree had moved from owning half of wealth to nearly three-quarters over this period. A good deal of this change is accounted for by the rising share of households with at least one member with a college degree. There were 26 million households where a member had a college degree in 1990, but 59 million in 2022. By contrast, the number of households with no BA was almost unchanged, rising from 68 million to 69 million.

VII. Mortality and Well-Being: Discussion

The results in this paper, on how people live and on how they die, should be seen in two different ways. The first is the documentation that the gaps between those with and without a college degree are not confined to one dimension of well-being, such as the mortality rates with which we began, but are pervasive across aspects of life that are important to people. Wherever we look, the more-educated group is faring better; sometimes

13. Board of Governors of the Federal Reserve System, "DFA: Distributional Financial Accounts," https://www.federalreserve.gov/releases/z1/dataviz/dfa/distribute/chart/, accessed July 23, 2023.

the college-educated are doing well and the noncollege-educated are losing ground, and sometimes both are seeing progress but the better-educated are seeing more.

The other way to look at the results is to use them to think about accounts of what is happening, about the *why* as well as the *what*. In our book on deaths of despair (Case and Deaton 2020), we suggest several mechanisms—the effects of globalization and automation without a European-style safety net and with an employer-based health insurance system that destroys good jobs, widens inequality, and lowers wages for less-skilled workers. Other rich countries do not finance health care this way. In our book we reference work that has documented an increase in corporate power relative to workers, the decline of unions, the spread of monopsony, and the decreased mobility of workers from less to more successful places. We also note again the evidence on some state legislatures passing business-written laws that harm workers.

Finally, we note the possibility that jobs are not always allocated by matching necessary or useful skills, but by the use of the BA as an arbitrary screen. We are encouraged by efforts by both public and private employers to remedy this; it is a low-cost policy that could have large benefits.

ACKNOWLEDGMENTS This is an extended and much revised version of NBER Working Paper 31236, May 2023; we are grateful to Jan Eberly, Jon Skinner, and Caroline Hoxby for comments on earlier versions. We gratefully acknowledge support from the National Institute of Aging through an award to the NBER, award number R01AG060104. We thank Andrew Foote for advice on education data and John Bound for clarification of the methods in Geronimus and others (2019).

References

Adler, Nancy E., Thomas Boyce, Margaret A. Chesney, Sheldon Cohen, Susan Folkman, Robert L. Kahn, and S. Leonard Syme. 1994. "Socioeconomic Status and Health: The Challenge of the Gradient." *American Psychologist* 49, no. 1: 15–24.

Arias, Elizabeth, Jiaquan Xu, Betzaida Tejada-Vera, Sherry L. Murphy, and Brigham Bastian. 2022. "U.S. State Life Tables, 2020." *National Vital Statistics Reports* 71, no. 2. Hyattsville, Md.: National Center for Health Statistics.

Arriaga, Eduardo E. 1984. "Measuring and Explaining the Change in Life Expectancies." *Demography* 21, no. 1: 83–95.

Autor, David, Arindrajit Dube, and Annie McGrew. 2023. "The Unexpected Compression: Competition at Work in the Low Wage Labor Market." Working Paper 31010. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31010.

Beltrán-Sánchez, Hiram, Samuel Preston, and Vladimir Canudas-Romo. 2008. "An Integrated Approach to Cause-of-Death Analysis: Cause-Deleted Life Tables and Decompositions of Life Expectancy." *Demographic Research* 19, article 35: 1323–50.

Blanchflower, David G., and Andrew J. Oswald. 2020. "Trends in Extreme Distress in the United States, 1993–2019." *American Journal of Public Health* 110, no. 10: 1538–44.

Burkhauser, Richard V., Kevin Corinth, James Elwell, and Jeff Larrimore. 2024. "Evaluating the Success of the War on Poverty since 1963 Using an Absolute Full-Income Poverty Measure." *Journal of Political Economy* 132, no. 1: 1–47.

Case, Anne, and Angus Deaton. 2017. "Mortality and Morbidity in the 21st Century." *Brookings Papers on Economic Activity*, Spring, 397–443.

Case, Anne, and Angus Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton, N.J.: Princeton University Press.

Case, Anne, and Angus Deaton. 2021. "Life Expectancy in Adulthood Is Falling for Those without a BA Degree, but as Educational Gaps Have Widened, Racial Gaps Have Narrowed." *Proceedings of the National Academy of Sciences* 118, no. 11: e2024777118.

Case, Anne, and Angus Deaton. 2022. "The Great Divide: Education, Despair, and Death." *Annual Review of Economics* 14: 1–21.

Case, Anne, Angela Fertig, and Christina Paxson. 2005. "The Lasting Impact of Childhood Health and Circumstance." *Journal of Health Economics* 24, no. 2: 365–89.

Cherlin, Andrew J. 2014. *Labor's Love Lost: The Rise and Fall of the Working-Class Family in America*. New York: Russell Sage Foundation.

Citro, Constance F., and Robert T. Michael, eds. 1995. *Measuring Poverty: A New Approach*. Washington: National Academies Press.

Crimmins, Eileen M., Samuel H. Preston, and Barney Cohen, eds. 2011. *Explaining Divergent Levels of Longevity in High-Income Countries.* Washington: National Academies Press.

Edin, Kathryn, Timothy Nelson, Andrew J. Cherlin, and Robert Francis. 2019. "The Tenuous Attachments of Working-Class Men." *Journal of Economic Perspectives* 33, no. 2: 211–28.

Farrell, Philip, and Victor R. Fuchs. 1982. "Schooling and Health: The Cigarette Connection." *Journal of Health Economics* 1, no. 3: 217–30.

Finkelstein, Amy, Casey McQuillan, Owen Zidar, and Eric Zwick. 2023. "The Health Wedge and Labor Market Inequality." *Brookings Papers on Economic Activity,* Spring, 425–75.

Geronimus, Arline T., John Bound, Timothy A. Waidman, Javier M. Rodriguez, and Brenden Timpe. 2019. "Weathering, Drugs, and Whack-a-Mole: Fundamental and Proximate Causes of Widening Educational Inequity in U.S. Life Expectancy by Sex and Race, 1990–2015." *Journal of Health and Social Behavior* 60, no. 2: 222–39.

Grumbach, Jacob M. 2022. *Laboratories against Democracy*: *How National Parties Transformed State Politics.* Princeton, N.J.: Princeton University Press.

Harris, Kathleen Mullan, Malay K. Majmundar, and Tara Becker, eds. 2021. *High and Rising Mortality Rates among Working-Age Adults.* Washington: National Academies Press.

Hayward, Mark D., and Mateo P. Farina. 2023. "Dynamic Changes in the Association between Education and Health in the United States." *Millbank Quarterly* 101, no. S1: 396–418.

Hendi, Arun S. 2017. "Trends in Education-Specific Life Expectancy, Data Quality and Shifting Educational Distributions: A Note on Recent Research." *Demography* 54, no. 3: 1203–13.

Ho, Jessica Y. 2017. "The Contribution of Drug Overdose to Educational Gradients in Life Expectancy in the United States, 1992–2011." *Demography* 54, no. 3: 1175–202.

Hummer, Robert A., and Joseph T. Lariscy. 2011. "Educational Attainment and Adult Mortality." In *International Handbook of Adult Mortality*, edited by Richard G. Rogers and Eileen M. Crimmins. Berlin: Springer Dordrecht.

Kitagawa, Evelyn Mae, and Philip M. Hauser. 1973. *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge, Mass.: Harvard University Press.

Krogstad, Jens Manuel, and Jynnah Radford. 2018. "Education Levels of U.S. Immigrants Are on the Rise." Pew Research Center, September 14. https://www.pewresearch.org/fact-tank/2018/09/14/education-levels-of-u-s-immigrants-are-on-the-rise/.

Lamba, Sneha, and Robert A. Moffitt. 2023. "The Rise in American Pain: The Importance of the Great Recession." Working Paper 31455. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31455.

Link, Bruce G., Mary E. Northridge, Jo C. Phelan, and Michael L. Ganz. 1998. "Social Epidemiology and the Fundamental Cause Concept: On the Structuring of Effective Cancer Screens by Socioeconomic Status." *Milbank Quarterly* 76, no. 3: 375–402.

Link, Bruce G., and Jo C. Phelan. 1995. "Social Conditions as Fundamental Causes of Disease." *Journal of Health and Social Behavior* (extra issue): 80–94.

Mackenbach, Johan P., José Rubio Valverde, Barbara Artnik, Matthias Bopp, Henrik Brønnum-Hansen, Patrick Deboosere, and others. 2018. "Trends in Health Inequalities in 27 European Countries." *Proceedings of the National Academy of Sciences* 115, no. 25: 6440–45.

Marmot, Michael. 2004. *The Status Syndrome: How Social Standing Affects Our Health and Longevity*. New York: Times Books.

Masters, Ryan K., Robert A. Hummer, and Daniel A. Powers. 2012. "Educational Differences in U.S. Adult Mortality: A Cohort Perspective." *American Sociological Review* 77, no. 4: 548–72.

McLanahan, Sara. 2004. "Diverging Destinies: How Children Are Faring under the Second Demographic Transition." *Demography* 41, no. 4: 607–27.

Meara, Ellen R., Seth Richards, and David M. Cutler. 2008. "The Gap Gets Bigger: Changes in Mortality and Life Expectancy, by Education, 1981–2000." *Health Affairs* 27, no. 2: 350–60.

Mehta, Neil K., Leah R. Abrams, and Mikko Myrskylä. 2020. "US Life Expectancy Stalls Due to Cardiovascular Disease, not Drug Deaths." *Proceedings of the National Academy of Sciences* 117, no. 13: 6998–7000.

Montez, Jennifer Karas, Jason Beckfield, Julene Kemp Cooney, Jacob M. Grumbach, Mark D. Hayward, Huseyin Zeyd Koytak, Steven H. Woolf, and Anna Zajacova. 2020. "US State Policies, Politics, and Life Expectancy." *Milbank Quarterly* 98, no. 3: 668–99.

Montez, Jennifer Karas, Robert A. Hummer, Mark D. Hayward, Hyeyoung Woo, and Richard G. Rogers. 2011. "Trends in the Educational Gradient of U.S. Adult Mortality from 1986 to 2006 by Race, Gender, and Age Group." *Research on Aging* 33, no. 2: 145–71.

Montez, Jennifer Karas, and Anne Zajacova. 2013a. "Explaining the Widening Education Gap in Mortality among U.S. White Women." *Journal of Health and Social Behavior* 54, no. 2: 166–82.

Montez, Jennifer Karas, and Anne Zajacova. 2013b. "Trends in Mortality Risk by Education Level and Cause of Death among US White Women from 1986 to 2006." *American Journal of Public Health* 103, no. 3: 473–79.

National Center for Health Statistics. 2022. *Instructions for Classification of Underlying and Multiple Causes of Death—Section 1—2022*. Washington: US Department of Health and Human Services. https://www.cdc.gov/nchs/nvss/manuals/2022/2a-2022.htm.

Novosad, Paul, Charlie Rafkin, and Sam Asher. 2022. "Mortality Change among Less Educated Americans." *American Economic Journal: Applied Economics* 14, no. 4: 1–34.

Olfson, Mark, Candace Cosgrove, Sean F. Altekruse, Melanie M. Wall, and Carlos Blanco. 2021. "Deaths of Despair: Adults at High Risk for Death by Suicide, Poisoning, or Chronic Liver Disease in the US." *Health Affairs* 40, no. 3: 505–12.

Olshansky, S. Jay, Toni Antonucci, Lisa Berkman, Robert H. Binstock, Axel Boersch-Supan, John T. Cacioppo, and others. 2012. "Differences in Life Expectancy Due to Race and Educational Differences Are Widening, and Many May Not Catch Up." *Health Affairs* 31, no. 8: 1803–13.

O'Rand, Angela M., and Scott M. Lynch. 2018. "Socioeconomic Status, Health, and Mortality in Aging Populations." In *Future Directions for the Demography of Aging: Proceedings of a Workshop*, edited by Malay K. Majmundar and Mark D. Hayward. Washington: National Academies Press.

Preston, Samuel H., and Paul Taubman. 1994. "Socioeconomic Differences in Adult Mortality and Health Status." In *Demography of Aging*, edited by Linda G. Martin and Samuel H. Preston. Washington: National Academies Press.

Rodrik, Dani. 1998. "Why Do More Open Economies Have Bigger Governments?" *Journal of Political Economy* 106, no. 5: 997–1032.

Rostron, Brian L., John L. Boies, and Elizabeth Arias. 2010. *Education Reporting and Classification on Death Certificates in the United States.* Vital and Health Statistics Series 2, no. 151. Hyattsville, Md.: National Center for Health Statistics.

Sasson, Isaac. 2016a. "Diverging Trends in Cause-Specific Mortality and Life Years Lost by Educational Attainment: Evidence from United States Vital Statistics Data, 1990–2010." *PLoS One* 11, e0163412.

Sasson, Isaac. 2016b. "Trends in Life Expectancy and Lifespan Variation by Educational Attainment: United States, 1990–2010." *Demography* 53, no. 2: 269–93.

Sasson, Isaac, and Mark D. Hayward. 2019. "Association between Educational Attainment and Causes of Death among White and Black US Adults, 2010–2017." *Journal of the American Medical Association* 322, no. 8: 756–63.

US Department of Health and Human Services. 2023. "New Surgeon General Advisory Raises Alarm about the Devastating Impact of the Epidemic of Loneliness and Isolation in the United States." Press release, May 3. https://www.hhs.gov/about/news/2023/05/03/new-surgeon-general-advisory-raises-alarm-about-devastating-impact-epidemic-loneliness-isolation-united-states.html.

Woolf, Steven H., and Laudan Aron, eds. 2013. *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. Washington: National Academies Press.

Comments and Discussion

COMMENT BY
CAROLINE HOXBY

STRENGTHS OF THE PAPER Case and Deaton thoroughly and transparently document that Americans who have a BA experience age-adjusted mortality at lower rates than those without such a degree.[1] They show, moreover, that this "mortality gap" has been growing over time. Adopting a novel approach, they mainly rely on death certificate data. These data have limitations, as discussed below, but they also have great advantages—namely, information on two variables that are central to the exercise. These variables are age at death and the proximate cause of death. While death certificate data are not available for all states in the years studied by Case and Deaton, it appears that when a state does make the individual-level data available, the data are comprehensive. Thus, sampling error is not an issue. Also, unlike Social Security death information, the death certificate data contain some demographic data.

I am persuaded by the authors' argument that mortality is an important indicator of a person's welfare and has several advantages over a measure such as wages. First, notwithstanding Horace's "Dulce et decorum est pro patria mori" (often translated as "It is sweet and fitting to die for one's country"), the vast majority of people agree that it is unambiguously negative to die unduly early or to die in suffering.[2] Second, mortality is a lifetime measure that can sum up many years and types of experience. In that sense,

1. The authors focus on two outcomes: age-adjusted mortality and life expectancy for 25-year-olds. For conciseness, I hereafter refer to these outcomes simply as mortality.

2. *Horace: Odes and Epodes*, trans. Niall Rudd (Cambridge, Mass.: Harvard University Press, 2004), 144–45.

Brookings Papers on Economic Activity, Fall 2023: 45–78

it resembles lifetime income rather than fluctuating wages. We also need not debate how to divide the earnings of salaried workers into wages and hours. Third, mortality is unusually comparable across time and space. There is no need to account for inflation or differences in the cost of living.

I am also persuaded that thought-provoking information is contained in a person's cause of death. As we know from their previous work, Case and Deaton (2017, 2020, 2022) are especially interested in "deaths of despair," in which they include deaths from drug overdoses, alcoholic liver disease, and suicide. The phrase is apt: these are often premature deaths associated not only with physical suffering but also with mental suffering. However, other causes of death are informative as well. Death from chronic lower respiratory disease may indicate a lifetime of tobacco smoking or exposure to air pollution. Death from diabetes hints at a lifetime of poor-quality foods, which can be cheaper than less-processed, fresher foods. Some of the evidence that may be unanticipated by readers suggests that part of the widening mortality gap may come from *breakthroughs* in medical treatment. Breast cancer is the most salient example. Breast cancer has traditionally been more prevalent in women who are better educated and more affluent.[3] Therefore, positive breakthroughs in breast cancer treatment are likely disproportionately to benefit more-educated women, widening the mortality gap. In short, cause of death may prompt us toward certain theories about mechanisms that lead to mortality.

Strikingly, the paper shows that the widening mortality gap is associated with causes that are becoming less prevalent for both BA holders and non-BA holders (cancer, cardiovascular disease), becoming more prevalent for both groups (deaths of despair, respiratory diseases, Alzheimer's disease), and becoming less prevalent among BA holders and more prevalent among non-BA holders (alcoholic liver disease, diabetes). This is remarkable: the widening mortality gap arises through all the possible channels. These findings suggest, at a minimum, that many mechanisms may contribute to the gap.

On a cautionary note, my review of the literature suggests that previous researchers have found that the cause of death information on death certificates is incorrect as much as half the time.[4] I return to this issue briefly below when discussing COVID-19.

3. Although some studies claim that the breast cancer–education correlation is due to more-educated women having later first births, a recent meta-analysis suggests that the evidence for this mechanism is less clear than commonly thought. See Løyland and others (2024).

4. There is a large body of research, often based on audits, showing that misreporting of cause of death is common. A good entry into the literature is McGivern and others (2017). My understanding is that age is much less likely to be inaccurate except in cases where the decedent does not die in a hospital, nursing home, or other health care facility.

A noteworthy strength of the paper is that many of the results can be immediately discerned from the figures. The tables mainly serve as confirmation.

DEATH CERTIFICATE DATA AND REVERSE CAUSALITY It is important to flag a potentially major reverse causality issue at the outset.

Since the educational attainment variable on death certificates is so important to the authors' exercise, it is crucial to know whether this variable is recorded accurately. With potential help from the next of kin, funeral directors usually fill in the answers to the questions on educational attainment, occupation, marital status, race, and ethnicity. Funeral directors do not ask for documentation such as college diplomas, college transcripts, or other evidence that a person has attained a BA.

This matters because of reverse causality. Suppose that the funeral director or the next of kin perceived the decedent to be intelligent, conscientious, articulate, planful, and capable of dealing with people who were college educated. Perhaps the decedent had an occupation that we would associate with a BA degree. Then the funeral director or next of kin might check the BA box on the form regardless of whether the decedent actually attained the degree. This action might seem appropriate to them, and their intentions would presumably be innocent. After all, the box-checking person would likely have no idea that the data might later be used to establish the empirical relationship between BA attainment and mortality.

However, inaccuracy of this type would matter a great deal because the decedent's BA designation would *not* be a cause of her acting intelligently, conscientiously, and so on. Rather, her behavior would be the cause of her BA classification. Since the same behavior could also presumably affect her mortality, it is crucial to know how death certificate data stand up to cross-validation from other, more authoritative, administrative sources. It would be unfortunate if reverse causality were an important explanation for the authors' results.

Rather surprisingly, the authors do not discuss the known tendency of educational attainment data from death certificates to overstate what people self-report through the Current Population Survey.[5] (The linked data set is known as the National Longitudinal Mortality Study.) I hesitate even to treat Current Population Survey data as a gold standard because people who are inclined to overstate their education on a survey may also overstate it to their family members. Ideally, we would like to have audits that rely on an authoritative administrative source such as the National Student

5. See Rostron, Boies, and Arias (2010); Rostron (2010); Feldman, Makuc, and Mussolino (1997); Sorlie and Johnson (1996); and Shai and Rosenwaike (1989).

Clearinghouse, which derives its individual-level longitudinal data from postsecondary institutions' records. Even validation using more aggregated data that institutions report to the US Department of Education (the Integrated Postsecondary Education Data System [IPEDS] and its predecessors) would be helpful.

Of course, what we want to know is not just whether BA attainment is overstated on death certificates. We want to know *for whom* it is overstated. Is it overstated for those whose behavior and environment are associated with low mortality (reverse causality)? Or is it overstated at random? The studies that rely on the National Longitudinal Mortality Study do not contain enough detail to answer this question well, but their findings provide a couple of hints. First, education is more likely to be overstated for people who are older when they die. Second, the studies find that much of the overstatement is among people who self-reported that they attended some high school or some college but who did not graduate with a high school degree or a BA, respectively. In other words, funeral directors and next of kin may use their discretion to "round up" to the next degree.

Later, I discuss the authors' within-cohort test for selection versus causality. That test relies on the assumption that people do not attain additional education after a certain age. Because reverse causality may affect death certificate data, that test is frail. In the funeral director and next of kin example above, the decedent would appear to have attained a BA late in life.

I discuss an additional problem with the relevant test below.

THE GENERAL PROBLEM OF CAUSALITY VERSUS SELECTION The alert reader may have noticed that, so far, I have avoided the language of causality but have written of associations, correlations, hints, suggestions, and the like. This restraint is because all of the facts and mechanisms described in the paper are consistent both with causal effects and selection. A causal effect would be one in which getting a BA degree literally causes people to change their behavior or environment in a way that reduces mortality. The most obvious example would be taking up an occupation that requires a BA degree because of licensing or a similar rule. If that occupation were physically safer, involved less exposure to pollution, or qualified people for more generous preventative health insurance, then the BA-to-mortality link would have a mechanism that could probably be demonstrated using statistical indices of on-the-job accidents, workplace air quality meters, or take-up of recommended tests (such as for colon cancer) that were paid for by health insurance.

However, the people who select into getting a BA degree may differ on numerous dimensions from non-BA holders. For instance, they may

discount the future less, as a matter of preference, and therefore invest more in both education and behaviors likely to prolong life. For instance, it is very plausible that people who discount the future less will find it preferable to refrain from smoking. Or, people who select into obtaining a BA may have higher native aptitude and thus be more likely to read medical instructions or compute nutritional content correctly. These would *not* be causal effects of the BA if the findings on the BA-mortality relationship would change substantially—or even disappear—if we were randomly to prevent some people from obtaining a BA that they would otherwise attain. We might also randomly treat some people with a degree (literally, force them through education and diploma receipt that they would otherwise not obtain), but this is a harder experiment to imagine.

The degree to which the BA-mortality relationship reflects causality or selection matters greatly because the policy implications differ. If the relationship is largely causal, society could improve mortality by inducing a larger share of people to attain a BA. Society could then not worry about addressing other possible mechanisms directly because the degree itself would generate the desired behaviors. The BA itself would cause smoking to fall. Anti-smoking laws and tobacco taxes would not be nearly as necessary.

It seems likely that some mechanisms are indeed causal as illustrated by the occupational example given above. Moreover, if selection into getting a BA had not changed over the period under study, one could not credibly construct a scenario in which selection accounted for much of the *change* in the mortality gap. That is, causal mechanisms would have to be at work if there were no changes in the nature of selection in BA attainment. Unfortunately for the causal case that the authors clearly wish to make (given the causal language that they consistently use), there have been very substantial changes in selection. Specifically, the share of each cohort obtaining a BA has risen greatly over time (shown below) even though the share of each cohort who are prepared for college has not improved in a parallel way. This makes it very unlikely that selection has not changed.

The authors are aware that changes in the nature of selection could pose a serious problem for causal interpretation of their findings. Indeed, the paper contains a short section that notes that selection could interact with mortality risk in ways that could be problematic and that cannot be ruled out except by making assumptions that cannot be verified with observable data. These issues, while known to the authors, were not covered in sufficient detail for audience members to grasp them fully. Thus, it might be helpful to show a few simple figures to illustrate the problem.

Fundamentally, we cannot observe a person's *latent mortality risk*, which is defined as the risk stemming from all factors that would exist in the absence of attaining a BA. It may help to think of a randomized trial in which some people are randomly forbidden to get a BA but are exactly the same people they would otherwise be. Factors that make latent mortality risk unobservable include preferences, aptitudes, genetics, home environment, and many behaviors that a statistician or econometrician cannot see or measure at all well.

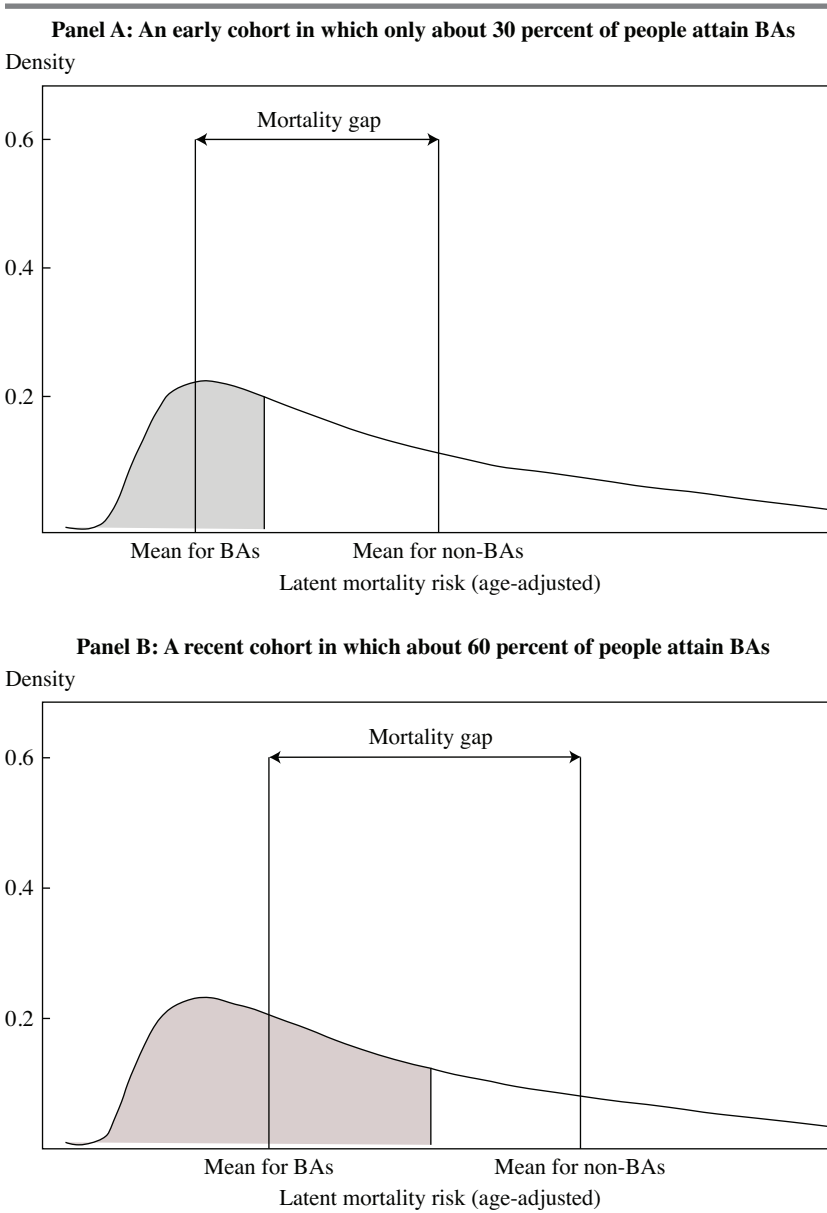
It is highly probable that the factors that affect latent mortality risk also affect a person's latent educational attainment—the education that a person would attain in the absence of any randomized intervention such as that described above. Any given factor might have a different effect on latent mortality risk than on latent attainment, so the two latent variables need not be highly correlated. However, to keep the figures in two rather than three dimensions, I assume that they are perfectly correlated. This is without loss of generality, but it makes the figures easier to interpret.

The figures are necessarily stylized since we do not know the distribution of latent mortality risk. Nevertheless, I have used Gompertz probability density functions since research suggests they fit *observed* mortality well, which, although not the same as *latent* mortality risk, probably reflects some of the shape of the latent risk (Juckett and Rosenberg 1993). I have also tried to stick fairly close to the facts, shown below, on the changing nature of selection into attaining a BA. For instance, I show the share of people with BA attainment rising by a realistic amount from early cohorts (about 30 percent) to recent cohorts (about 60 percent). It should be understood, however, that this is a demonstration of the importance of knowing the distribution of latent mortality. It is not an empirical analysis.

Figure 1 illustrates a situation in which the changing nature of selection into the BA is only a moderate problem. In panel A, about a third of an early cohort, shown in the shaded part of the probability density function, get a BA degree. The BA holders are drawn from the lowest part of the mortality risk distribution. Thus, BA holders have lower average mortality risk than the average risk of non-BA holders. The latent mortality gap is the distance between the non-BA holders' average risk and the BA holders' average risk.

Panel B represents a recent cohort in which BA attainment is less selective. That is, a larger share of the cohort, shown as two-thirds of the distribution, get a BA. Again, the mortality gap is shown as the distance between the non-BA holders' average risk and the BA holders' average risk. The mortality gap has risen by about 28 percent due entirely to the changing nature of

Figure 1. Mortality Gap Derived from a Latent Mortality Risk Distribution with a Low Peak



Source: Author's illustration.

Note: The figure is a stylized representation in which people who attain BAs have lower latent mortality than people who do not attain BAs. The mortality gap is defined as the average latent mortality risk among non-BAs minus the average latent mortality risk among BA holders. The distribution is based on the shape of a Gompertz distribution with $\alpha = 1.3$, $\beta = 1.2$, and $\gamma = 0.7$.

selection, with BA attainment having no causal effect on mortality. (The exact percentage increase does not matter.) The mortality gap, in the case illustrated, rises moderately purely through selection because the marginal "switchers" into the BA group have sufficiently low mortality risk that, although their joining the BA group raises the average risk in both groups, it raises it more in the non-BA group than in the BA group.

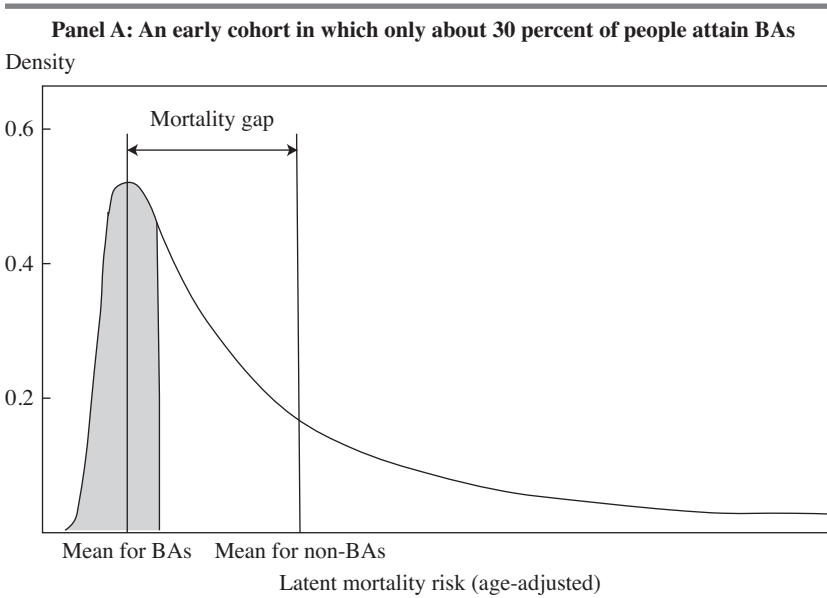
Fgure 2, panel A, shows an early cohort with a Gompertz-type density that is more strongly peaked in the lower range of mortality risk. (By more strongly peaked, I mean that α is lower while β and γ are the same as in figure 1.) Again, about a third of the early cohort get a BA degree. They are in the shaded part of the distribution and have very low average mortality risk owing to the shape of density function. Average latent mortality risk among non-BA holders is substantially higher. Notice that the non-BA holders include both some very low-risk people and a long tail of high-risk people. As in the previous figure, the mortality gap is the difference in average latent mortality risk between the non-BA holders and the BA holders.

Finally, figure 2, panel B, represents a recent cohort with the more strongly peaked distribution of mortality risk. As in figure 1, panel B, about two-thirds of the cohort get a BA because, in recent years, attainment has become less selective. Compared to that of the early cohort (panel A), the mortality gap has risen sharply. Specifically, the mortality gap has risen by about 60 percent due entirely to the changing nature of selection, with BA attainment having no causal effect on mortality. As in the previous example, the switchers into the BA group raise the average risk of both the BA holder and the non-BA holder groups. However, since the density is so peaked in the lower range of risk, the BA holders' risk does not rise nearly as much as the non-BA holders' risk, the latter of which reflects the distribution's long tail.

It should now be clear that the *shape* of the latent mortality distribution matters a great deal. But this is a shape that we cannot observe because the latent risk is, well, *latent*. Thus, both of the previous examples are plausible, and it is impossible to determine the true role of selection in causing the mortality gap to expand.

Since the latent distribution's shape matters, it is possible to devise examples in which selection has no effect on the mortality gap because the switchers generate an equal rise in the mortality risk of both the BA and non-BA groups. This type of example is one emphasized by the authors. It is even possible to devise examples in which selection lowers the mortality gap because the shape of distribution is such that switchers generate only a small rise in risk among non-BA holders but generate a large rise in risk among BA holders. However, this type of example is not worth illustrating

Figure 2. Mortality Gap Derived from a Latent Mortality Risk Distribution with a High Peak



Source: Author's illustration.

Note: The figure is a stylized representation in which people who attain BAs have lower latent mortality than people who do not attain BAs. The mortality gap is defined as the average latent mortality risk among non-BAs minus the average latent mortality risk among BA holders. The distribution is based on the shape of a Gompertz distribution with $\alpha = 0.8$, $\beta = 1.2$, and $\gamma = 0.7$.

here because it does not seem pertinent to the paper under discussion. Moreover, it is easiest to create such examples with distributions that have long left tails and peak density on the right. While I cannot miraculously observe latent densities, it is doubtful whether distributions with such shapes are relevant. This is owing to the aforementioned tendency of Gompertz-shaped distributions to fit observed mortality data best.[6]

Even though I kept my examples simple, they make it clear that there are no easy ways to quantify the degree to which the observed increase in the mortality gap reflects causal effects versus selection. I discuss possible quasi experiments below.

A BRIEF REVIEW OF THE CHANGING NATURE OF SELECTION INTO BA ATTAINMENT

It is worthwhile showing just a few obvious pieces of evidence on the changing nature of selection into BA attainment.

In a nutshell, a larger share of each high school graduating cohort has obtained a BA. This is despite later cohorts being apparently no more prepared than earlier cohorts. This suggests that BA granting has become a less selective and probably less challenging process over time. This is not surprising because many of the additional seats that have been added in postsecondary education are in colleges that have always been nonselective or barely selective.[7] That is, seats have been disproportionately added at schools that anyone with a high school degree or General Educational Development (GED) can attend. Seats have also been disproportionately added at institutions that are for-profit, online, or both.

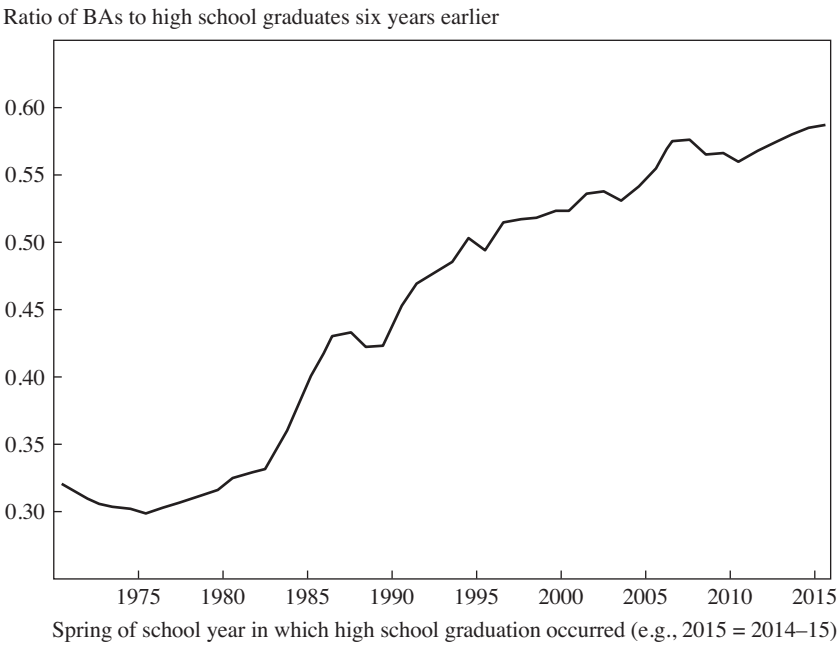
Figure 3 shows that the ratio of the number of BAs conferred to the number of high school graduates doubled from 30 percent among 1975 high school graduates to 59 percent among their 2015 counterparts.[8]

6. Distributions that compete with Gompertz are the Weibull and lognormal distributions. These have similar shapes to the Gompertz distributions and do not exhibit long left tails and density peaks in the high-risk range. See Juckett and Rosenberg (1993).

7. Author's calculations based on IPEDS data up through 2022 (the most recent year). For a summary of similar results that are not quite so recent, see Baum, Kurose, and McPherson (2013).

8. High school graduates in 2015 are the most recent for whom such numbers are available. It is conventional in education policy research to allow a lag of six years between high school graduation and the attainment of a BA. This is known as completion within 150 percent of time, and statistics on on-time completion tend to be recorded with this lag. See online documentation for IPEDS. The *Digest* data used to construct figure 3 are derived from the school-level data in IPEDS and the Common Core of Data, both of which are provided by the National Center for Education Statistics. One can make more-detailed calculations using IPEDS institutional data on completions by age for 150 percent of time and 200 percent of time. Such calculations produce similar patterns as figure 3 shows. The complexities involved in making such calculations could not be properly described in a short discussion.

Figure 3. Ratio of BAs to High School Graduates Six Years Earlier: High School Graduates from 1970 to 2015



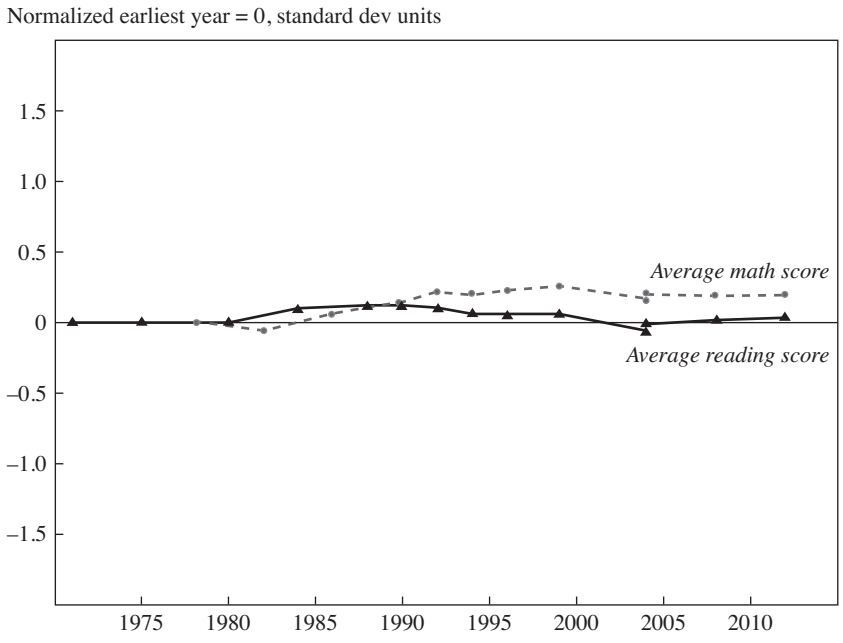
Source: Author's calculations based on NCES, *Digest of Education Statistics*: 2021, tables 219.10 and 322.10; 2018, table 322.10; 2013, table 318.10; 1995, tables 98 and 236.

Note: Later *Digest* published numbers are used in preference to earlier published ones, which are more likely to have been revised.

We can get a sense of the changing nature of selection when we compare the doubling of the share attaining a BA to the lack of improvement in precollege achievement. Figure 4 shows the results of high school seniors (17 years old) on the National Assessment of Education Progress (NAEP) long-term trend tests in mathematics and reading. These are tests given to nationally representative samples of students. The long-term trend tests are deliberately designed to facilitate comparisons over decades. Figure 4 shows the results in standard deviation units where the earliest year's results are normalized to zero, both for reading and mathematics. This is a conventional way to represent scores that would otherwise be on an unfamiliar scale that readers would find hard to interpret.[9]

9. See Beaton and Chromy (2010). Page 52 is especially relevant.

Figure 4. NAEP Math and Reading Scores among 17-Year-Olds (High School Seniors): High School Students from 1970 to 2012



Source: Author's computations based on reports derived from the NAEP long-term trend data reports for 17-year-olds, https://www.nationsreportcard.gov/ndecore/xplore/ltt.

Note: Scale scores are normalized so that the earliest year shown has its score equated to zero. The scores are shown in standard deviation units, and the standard deviations are based on Beaton and Chromy (2010).

If US students' achievement were improving relative to the earliest years in which the tests were given, then we would expect a rise in scores by at least one standard deviation between the 1971 high school senior cohort (the earliest) and the 2012 cohort (the latest). These are, after all, forty-two cohorts who cover the dramatic growth in BA attainment, shown in figure 3. It is not only the *average* high school senior's NAEP scores that have hardly budged over four decades. The *distribution* of scores (not shown here) has also not changed much. Based on the latest 2019 "main NAEP" tests of high school seniors, only about 24 percent could fairly confidently be predicted to be "college-ready" in mathematics, according to the ACT's empirically based standard. A similar percentage are "college-ready" in reading. In short, only about a quarter of high school seniors are well prepared to thrive in

college.[10] Yet, in recent cohorts, about 60 percent attain a BA. Selection into the BA has apparently changed.

Other evidence that selection into the BA has changed comes from the National Center for Education Statistics high school longitudinal studies of the high school graduating classes of 1972, 1982, 1992, and 2004.[11] These studies contain mathematics tests taken by nearly all the participants, and the tests are designed to be comparable over all the graduating classes.[12] The study participants are followed for at least eight years after their senior year of high school.

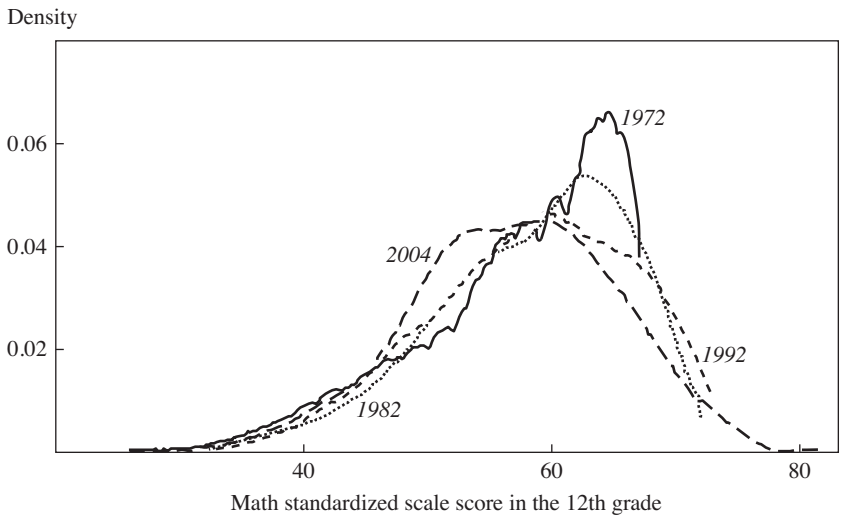
Figure 5 shows that the distribution of high school mathematics scores among BA holders has been shifting downward from the 1972 graduating cohort to the 1982 cohort to the 1992 cohort to the 2004 cohort. The mean, median, and mode are all shifting downward. Moreover, the distribution

10. There is a strong psychometric relationship between the long-term trend NAEP and main NAEP, the latter of which is designed to be more flexible across years. See Beaton and Chromy (2010). See Xi and others (2020), pages 10–11 for conversions between the main NAEP and college-readiness. In mathematics, one can be about 80 percent confident that students who meet the Proficiency standard (score of 176) on the main NAEP are college-ready (a very similar score of 180). The source for the 2019 percent Proficient and Above mathematics number is *Digest of Education Statistics*: 2022, table 222.12. In reading, the college-readiness standard (a score of 324) lies midway between the Proficient standard (a score of 302) and Advanced standard (a score of 346) on the main NAEP. Since only about half (about 15.5 percent) of the Proficient students are college-ready while all 6 percent of the Advanced students are college-ready, the total percent of the students who are college-ready in reading is approximately 21.5 percent. The source for the 2019 percent Proficient and Advanced reading numbers is *Digest of Education*: 2022, table 221.12.

11. The studies are the National Longitudinal Study of the Class of 1972 (NCES 1981), High School and Beyond (class of 1982), the National Education Longitudinal Study (class of 1992), and the Education Longitudinal Study (class of 2004). Unfortunately, the most recent study (class of 2013) has not yet been followed up long enough for us to ascertain who will and will not earn a BA. For a description of all the studies and their design, see National Center for Education Statistics (NCES), "Secondary Longitudinal Studies Program," https://nces.ed.gov/surveys/slsp/. The data sources are NCES, *National Longitudinal Study of 1972: Base Year (1972) through Fourth Follow-Up (1979)*, electronic data (1981); *High School and Beyond Fourth Follow-up (Sophomore Cohort) HS&B 1992,* electronic data from NCES 95305 (1995); *National Education Longitudinal Survey of 1988 (NELS88) Base Year through Fourth Follow-up*, electronic data from NCES 2003-348 (2003); *ELS: 2002 Base Year to Third Follow-up Postsecondary Transcripts*, electronic data in NCES 2015-314 (2015); *National Longitudinal Study of 1972: Base Year (1972) through Fourth Follow-Up (1979)*, electronic data (1981).

12. Unfortunately, only mathematics tests are available for all of the cohorts. However, mathematics scores are highly correlated with reading, science, and social studies scores for the cohorts that have the full array of scores available. The sample in each study is designed to be nationally representative when the appropriate sample weights are used.

Figure 5. Distribution of Math Scores among High School Seniors: High School Seniors in the Graduating Classes of 1972, 1982, 1992, and 2004



Source: Author's calculations.

Note: Density plots of 12th-grade math scores of participants in secondary school longitudinal surveys implemented by NCES for 1981, 1995, 2003, and 2015.

of college-incoming scores has been widening, mainly because additional density has been added to low range of scores. Simply put, students whose scores would not have led them to BAs in earlier years are, in recent years, attaining BAs. This is an indicator that selection into the BA has changed.

Over time, much of the growth in BAs has come from schools that have never been selective in the sense that any student can enroll who has a high school degree or passing score on the GED test. While some of the growth is attributable to publicly controlled colleges, much of the recent growth is attributable to for-profit schools, a good share of which are wholly or partly online (Hoxby 2018a). There is controversy about whether these schools provide rigorous educational experiences. At these schools, a large share of students who are enrolled in BA programs drop out long before attaining a BA. However, the students who do persist, even if they are not stellar academically, may have traits that are valuable for reducing mortality risk. For instance, the students who attain BAs in these unpropitious environments may have high long-term orientation, grit, motivation, or support from their families. This is a speculation based on my analysis that shows that only students who persist over five or more years realize returns to this type of postsecondary education (Hoxby 2018b).

While on the topic of nonselective, for-profit, and online schools, it should be noted that students at these institutions are, on average, in their mid-thirties, not in their late teens or early twenties. The average age is 35 among students in schools that are at least partially online, and numerous students are in their forties (Hoxby 2018b). Such students often say that they are seeking BAs because they are "getting their life together" or realize that they made poor educational decisions when they were younger. These facts matter because the authors' main test of whether selection matters depends on there being little or no actual growth in BA attainment within a cohort over time, but schools that serve older students represent the fastest growing sector of postsecondary education and the older students who do attain BAs may be especially capable.

MIGHT NATURAL OR QUASI EXPERIMENTS IDENTIFY THE CAUSAL EFFECTS OF BA ATTAINMENT ON MORTALITY? While preparing to discuss the authors' paper, I wracked my brain in an attempt to think of a natural or quasi experiment that could credibly identify the causal effects of a BA degree on mortality among Americans. I did this for two reasons. The first is simply that I enjoy being constructive in this way. The second is that the exercise is a good way to sharpen one's thinking on the sources of variation in an outcome. If one cannot think of any exogenous or arbitrary sources of variation in an outcome that could account for the observed scale of the variation in the outcome, then perhaps there really is not much exogenous variation. Some phenomena are generated by interactions that are too complex or subtle to be reduced to an effect that can be described simply, such as the effect of having a BA. This does not mean that the phenomena are not real. For instance, many people believe that love is a real phenomenon and that people who experience more love have better outcomes. However, it would seem almost absurd to argue that if person A could just induce person B to love her, person A would have better outcomes. This would be the stuff of love elixirs from Jacobean drama.

Returning to the problem of BA attainment, I considered the numerous natural, policy, or quasi experiments that credibly prompt some people to get a BA when they would not otherwise do so. Most often, these are scholarships or other inducements to attain a BA degree. Relevant studies occasionally rely on actual randomization but more often rely on empirical designs such as a regression discontinuity in the eligibility for the scholarship. Such studies credibly identify the causal effects of BA receipt on early career earnings, unemployment, and many more outcomes. However, these studies typically do not lend themselves to mortality as an outcome because it is so uncommon among the relatively young that almost any study that

does not depend on a very large-scale experiment will fail for reasons of statistical power.

Angrist (1990), in a well-known paper, used a person's draft lottery number as an instrument for serving in the Vietnam War. This is a quasi experiment on such a large scale that statistical power is not an issue. Moreover, veterans were eligible for generous college financial aid after returning to the United States. So one might surmise that draft lottery numbers were a credible instrument for attaining a BA and thus for obtaining estimates of the causal effects of a BA on mortality. Indeed, Vietnam era people are sufficiently aged at present that they are at reasonable risk of mortality. However, as Angrist himself would almost certainly argue, the draft lottery affected outcomes other than educational attainment—most importantly, service in Vietnam. Since such service might easily affect mortality through exposure to war-related disabilities, trauma, exposure to Agent Orange, and a myriad of other phenomena, it would be nigh impossible to disentangle the role of BA attainment on mortality. Quasi experiments along these lines, including those that rely on various GI Bill benefits, often run into such difficulties, although the difficulties can sometimes be overcome.

Another quasi experiment that is seemingly close to what the authors want to turn on and off is the Chinese Cultural Revolution, during which many people who would otherwise have obtained a university degree were forcibly sent to rural China and forbidden from pursuing higher education. One might think that exposure to the Cultural Revolution was quasi-random. After all, some people were born in a cohort that was less exposed. Others were born in a proximate cohort that was fully exposed. Here, we have an experiment of incredibly large scale in which not merely the university diploma is turned on and off. Many of the mechanisms that the authors describe as influenced by BA receipt are potentially affected as well. The problem is that the Cultural Revolution had dramatic general equilibrium effects. It greatly changed universities (depriving them of skilled faculty), generated chaos in the economy, and affected some people and regions far more than others ("conservative" people were more likely targets, and some areas experienced much more violence).

A final quasi experiment, one that may hold some promise for exercises like the authors', is relying on differences among US states in the timing and level of their support for public universities. Increases in such support appear to induce more students to complete BAs (Bound, Lovenheim, and Turner 2010). Since death certificates include specific locational data as well as age data, one might gain traction on causality versus selection using state-by-time differences in colleges' funding and seats. A researcher would

need to argue that the timing of sharp funding differences is quasi-random within proximate cohorts and is unrelated to other coincident phenomena such as local economic downturns. The study closest in spirit is Fletcher and Noghanibehambari (2024), although they use college expansions, which have been shown to have problematic associations with variables that reflect an area's improving population and/or improving economy.

Summing up, the exercise of thinking through numerous quasi experiments did not impress me with the idea that BA attainment has been affected by exogenous forces of sufficient scale and impact to account causally for all—or even the vast majority—of the observed changes in the relationship between mortality and BA completion. I would therefore counsel more reticence regarding language and arguments that explicitly or implicitly make claims for causal effects, even if causal effects account for a substantial share of the facts described. Descriptive evidence makes important contributions to economics because it arms us with facts that we must work to explain. However, a conflation between descriptive evidence and credibly causal evidence—such as often occurs in nonexperimental health research—is not especially helpful to refining economists' logical skills.

HAVING A BA AND REMOTE WORK DURING COVID-19 The authors are careful to show the mortality gap with and without deaths attributed to COVID-19. Such evidence is helpful, and I was grateful for it when reading the paper. However, I find the COVID-19 evidence to be somewhat unconvincing because many deaths that were related to COVID-19 did not record the virus as the proximate cause of death. This has been shown convincingly in studies of excess mortality (Paglino and others 2024). Thus, removing the deaths that were formally attributed to COVID-19 does not solve the problem that mortality gaps expanded in a way that were highly anomalous during the pandemic. I find the chasmal mortality gaps in that period to be uninformative.

Moreover, when the authors argue, albeit with caution, that the pandemic-related changes in the mortality gap are useful, they unintentionally undermine their argument that selection is unimportant. Selection into COVID-19 exposure was involuntary for many people whose *existing* jobs made it difficult to work remotely or telecommute, in the language used by the Bureau of Labor Statistics (BLS). The pandemic was a temporary and unforeseen shock to the mortality risk associated with being in proximity to other people. It was not a shock to BA attainment. It also did not trigger a permanent change in mortality risks that might be *caused* by attaining a BA—such as a BA being a condition for a license in occupations that are permanently associated with low health risks or better health insurance. Rather, people

Table 1. Percentage of Employed People Who Worked Remotely during COVID-19 and in Occupations Classified as Suitable for Remote Work, by Educational Attainment

Educational attainment	*Teleworking during COVID-19, by educational attainment*	*In a suitable job for teleworking*
Less than a high school diploma	3.3	10.2
High school graduate, no college	8.8	25.8
Some college, associate's degree	16.9	40.3
Bachelor's degree only	40.6	63.4
Advanced degree	54.4	71.3

Source: Data from Dey and others (2021).

who already lacked BAs were disproportionately likely to be incumbents in jobs that were unsuitable for temporary remote work. This is not an argument for the power of the causal mechanisms that could plausibly have been the source of the growth in the mortality gap *over decades*. If grocery store cashiers or meat-processing workers had experienced helicopter drops of BA diplomas, their COVID-19-related exposure risks would not have decreased precipitously because they would have, say, suddenly and voluntarily adopted healthier behaviors. Therefore, the sharp and dramatic mortality increases among non-BA holders is not a causal effect of their lacking BAs.

In table 1, I show results from a recent BLS study that shows that non-BA holders were much less likely to work remotely during the height of the pandemic. The second column in the table shows that a primary explanation for this phenomenon is that their existing jobs were less suitable for remote work. The BLS study does not attempt to argue that a lack of attainment *caused* the non-BA holders voluntarily to adopt less healthy behaviors or that having a BA would have quickly switched them to healthier environments.

ADDITIONAL THOUGHTS I do not see why, based on causal logic, one would prefer a binary BA/non-BA measure to more continuous measures of cognition, achievement, or attainment. Nearly all of the causal arguments made by the authors are inherently *continuous*, not discrete at the margin of obtaining a BA. For instance, if improved health behaviors are caused by increases in knowledge, such improvements would surely be continuous in educational attainment, not affected discretely by the receipt of a BA diploma.

Once a state starts asking about educational attainment on its death certificates, its categories are several, not just non-BA versus BA. For instance,

a Pennsylvania death certificate provides multiple categories of attainment: 8th grade or less, 9th through 12th grades with no diploma, high school graduate or GED, some college but no degree, associate's degree, bachelor's degree, master's degree, doctorate or professional degree. Since the causal arguments for the effects of educational attainment on mortality are continuous, there would seem to be little reason for the authors to rely exclusively on the binary BA measure. The literature on signaling has long associated certain degrees with being signals of unobserved aptitudes. Classic signaling is an expression of equilibrium *selection on unobserved traits*. While I am certainly not one to argue that most education is a signal rather than an investment in human capital, I see no reason to focus on the discrete BA measure. By using more continuous measures of attainment, the authors might allay some concerns about selection versus causality.

It would be useful to distinguish between changes over time that are due to behaviors that people themselves at least partially control (diet, substance abuse) and changes that could not possibly be controlled by an individual (medical advances in heart surgery or cancer treatment). The distinction is important because the latter causal mechanisms can only run through processes that are observable and thus testable. For instance, suppose a person has a cancer for which there is a medical breakthrough. It might be that BA holders get the new treatment first or attend their therapy sessions more regularly. However, the BA holders do not determine the timing of the breakthrough: earlier cohorts might have died even if they were vigilant about preventative medicine and diagnosis. Furthermore, medical data would allow us to observe that BA holders were indeed obtaining the breakthrough procedures. We would also likely be able to link the BA holders to what was allowed under their health insurance. Such intermediate evidence on mechanisms can help support arguments for causality.

In contrast, individuals' actual diets are largely under their control and mysterious to econometricians—sometimes even to their fellow household members. Even Nielsen households can strategically omit to record their consumption of junk food or alcohol. Thus, we have no real hope of getting accurate, administrative data on dietary mechanisms that would be analogous to the data we could obtain on cancer treatment. As a result, the problem of selection is far less remediable for certain proposal channels—such as diet—of causal BA effects.

Summing up, I derived a lot of benefit from this paper for all of the reasons stated in the first section. It is extremely thorough and contains many striking results, presented coherently. However, my own interpretation is much more cautious, with regard to causality, than that of the authors.

REFERENCES FOR THE HOXBY COMMENT

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80, no. 3: 313–36. http://www.jstor.org/stable/2006669.

Baum, Sandy, Charles Kurose, and Michael McPherson. 2013. "An Overview of American Higher Education." *Future of Children* 23, no. 1: 17–39.

Beaton, Albert E., and James R. Chromy. 2010. *NAEP Trends: Main NAEP vs. Long-Term Trend*. San Mateo, Calif.: American Institutes for Research. https://www.air.org/sites/default/files/2021-06/NAEP_Trends_12-6-10_0.pdf.

Bound, John, Michael F. Lovenheim, and Sarah Turner. 2010. "Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources." *American Economic Journal: Applied Economics* 2, no. 3: 129–57.

Case, Anne, and Angus Deaton. 2017. "Mortality and Morbidity in the 21st Century." *Brookings Papers on Economic Activity*, Spring, 397–476.

Case, Anne, and Angus Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton, N.J.: Princeton University Press.

Case, Anne, and Angus Deaton. 2022. "The Great Divide: Education, Despair, and Death." *Annual Review of Economics* 14: 1–21.

Dey, Matthew, Harley Frazis, David S. Piccone Jr., and Mark A. Loewenstein. 2021. "Teleworking and Lost Work during the Pandemic: New Evidence from the CPS." *Monthly Labor Review*, Bureau of Labor Statistics. July. https://www.bls.gov/opub/mlr/2021/article/teleworking-and-lost-work-during-the-pandemic-new-evidence-from-the-cps.htm.

Feldman, Jacob J., Diane M. Makuc, and Michael E. Mussolino. 1997. "Validity of Education and Age as Reported on Death Certificates." In *1996 Proceedings of the Social Statistics Section*. Alexandria, Va.: American Statistical Association.

Fletcher, Jason, and Hamid Noghanibehambari. 2024. "The Effects of Education on Mortality: Evidence Using College Expansions." *Health Economics* 33, no. 3: 541–75. https://doi.org/10.1002/hec.4787.

Hoxby, Caroline M. 2018a. "Online Postsecondary Education and Labor Productivity." In *Education, Skills, and Technical Change: Implications for Future US GDP Growth*, eds. Charles R. Hulten and Valerie A. Ramey. Chicago: University of Chicago Press.

Hoxby, Caroline M. 2018b. "Online Postsecondary Education and the Higher Education Tax Benefits: An Analysis with Implications for Tax Administration." *Tax Policy and the Economy* 32, no. 1: 45–106.

Juckett, David A., and Barnett Rosenberg. 1993. "Comparison of the Gompertz and Weibull Functions as Descriptors for Human Mortality Distributions and Their Intersections." *Mechanisms of Ageing and Development* 69, no. 1–2: 1–31.

Løyland, Borghild, Ida Hellum Sandbekken, Ellen Karine Grov, and Inger Utne. 2024. "Causes and Risk Factors of Breast Cancer, What Do We Know for Sure? An Evidence Synthesis of Systematic Reviews and Meta-Analyses." *Cancers (Basel)* 16, no. 8: 1583.

McGivern, Lauri, Leanne Shulman, Jan K. Carney, Steven Shapiro, and Elizabeth Bundock. 2017. "Death Certification Errors and the Effect on Mortality Statistics." *Public Health Reports* 132, no. 6: 669–75.

Paglino, Eugenio, Dielle J. Lundberg, Elizabeth Wrigley-Field, Zhenwei Zhou, Joe A. Wasserman, Rafeya Raquib, and others. 2024. "Excess Natural-Cause Mortality in US Counties and Its Association with Reported COVID-19 Deaths." *Proceedings of the National Academy of Sciences* 121, no. 6: e2313661121.

Rostron, Brian L. 2010. "Socioeconomic Differences in Education Reporting and Their Effect on Estimates of Life Expectancy by Educational Attainment in the U.S." Paper prepared for the Population Association of America 2010 Annual Meeting, Dallas, Texas, April 15–17. https://paa2010.populationassociation.org/papers/100300.

Rostron, Brian L., John L. Boies, and Elizabeth Arias. 2010. "Education Reporting and Classification on Death Certificates in the United States." *Vital and Health Statistics* Series 2, no. 151. Hyattsville, Md.: National Center for Health Statistics.

Shai, Donna and Ira Rosenwaike. 1989. "Errors in Reporting Education on the Death Certificate: Some Findings for Older Male Decedents from New York State and Utah." *American Journal of Epidemiology* 130, no. 1: 188–92.

Sorlie, Paul D., and Norman J. Johnson. 1996. "Validity of Education Information on the Death Certificate." *Epidemiology* 7, no. 4: 437–39.

Xi, Nuo, Mei-Jang Lin, Laura Jerry, David Freund, and Helena Jia. 2020. "NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between NAEP and ACT Assessments in Reading and Mathematics." Washington: National Center for Education Statistics.

COMMENT BY

JONATHAN SKINNER[1] The association between education and mortality has been well understood for more than a half century. In a remarkable study, Kitagawa and Hauser (1968, 1973) and their team linked 340,000 death records from 1960 to the recently conducted 1960 US Census to measure the education-mortality gradient at the national level. For people with fewer than eight years of education (which at the time comprised nearly a quarter of the population), they found 48 percent higher midlife (age 25–64) mortality among white men and 68 percent among white women, compared to those with some college. While these mortality gaps in 1960 were substantial, they have grown much larger since then. By 1986, the midlife mortality ratio for college graduates relative to those without a high school degree had risen to 171 percent for white men and 88 percent for white women

1. I am grateful to Christopher Foote and Ellen Meara for helpful suggestions.

(Pappas and others 1993).[2] The corresponding rate for Black men was 123 percent, and for Black women 182 percent.

As Case and Deaton have documented in this paper, the gap is not just "rising" as Meara, Richards, and Cutler (2008) documented during the 1980s and 1990s, but "exploding."[3] As they show, for people age 25–84, the mortality gap between noncollege graduates and college graduates has risen from 211 per 100,000 in 1992 to 643 per 100,000 in 2021. The corresponding midlife (age 25–64) mortality rate by 2019 for noncollege graduates was four times the rate for college graduates (Foote and others 2024). As Case and Deaton (2021) have shown, the difference in life expectancy between college and noncollege graduates exceeds the gap between Black and non-Hispanic white populations and between Hispanic and non-Hispanic white populations.

One may be concerned with these comparisons given selection effects; the fraction of people who are college graduates has been rising since 1992, while the fraction of those who did not complete high school has been declining rapidly. Case and Deaton argue persuasively that selection is not the likely explanation for their results, although there is some question about whether "noncollege graduates" masks heterogeneity within this group. While Leive and Ruhm (2021) show a widening educational gradient in mortality across all percentiles of the education distribution, Novosad, Rafkin, and Asher (2022) argue that most of the decline in mortality for noncollege graduates is the consequence of a steeply increasing gradient at the very bottom of the education distribution.

THE COVID-19 PANDEMIC Unlike previous studies by Case and Deaton, which focused on midlife mortality and later the average number of years lived from 25–75, this paper considers the average number of years lived between age 25–84 (so the theoretical maximum is sixty years). This lengthier horizon dilutes the impact of deaths of despair somewhat because they are only a small fraction of total deaths (although weighted more heavily because of the greater loss in life-years). But considering these older populations better captures the differential impact of the COVID-19 pandemic, which disproportionately affected older people. As Case and Deaton show, the pandemic caused a dramatic increase in the educational

2. It is difficult to line up measures of "high" and "low" education over time as rates of high school and college graduation have risen since 1960; these selection issues are discussed below.

3. Focusing on life expectancy, Meara, Richards, and Cutler (2008) found life expectancy from the late 1980s to the late 1990s grew by 1.4 years for people with high levels of education compared to just 0.5 years for those with lower levels of education.

gradient, a result that has also been found for the income gradient (Schwandt and others 2022).

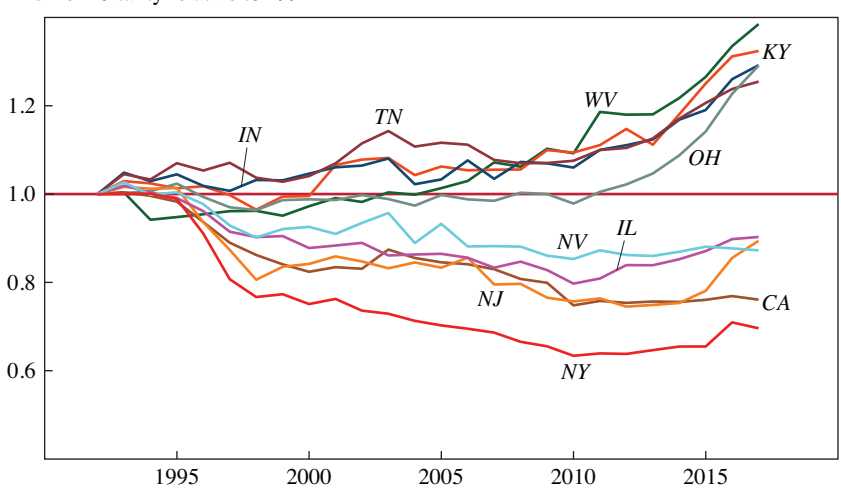
We might expect the COVID-19 educational gradient to subside somewhat simply because the number of deaths reported where COVID-19 was the underlying cause or a contributing cause declined from about 463,000 in 2021 to just 61,000 through the first week of November 2023.[4] Yet, in many ways Case and Deaton's most striking finding is the increase in non-COVID-19 mortality, particularly from deaths of despair, which would be less likely to be misdiagnosed or caused directly by COVID-19. These only accelerated during the pandemic, with alcohol-related deaths for those without a BA rising by ten per 100,000 between 2019 and 2021, more than the entire increase of seven per 100,000 during the twenty-seven years prior to the pandemic. There are fewer signs that these non-COVID-19 shifts in mortality are reverting to pre-COVID-19 levels; opioid deaths continued to exceed 100,000 in 2022 (NCHS 2023).

WHAT CAN EXPLAIN THE DIVERGENCE IN MORTALITY BY EDUCATION? There is substantial literature on the higher life expectancy associated with education, but understanding why such differences exist is less well understood (Cutler and Lleras-Muney 2010). It could be that education per se—skills and reasoning learned in the classroom—could lead to greater longevity, but the empirical evidence supports at best just modest effects of an exogenous increase in schooling on longevity and health (Galama, Lleras-Muney, and van Kippersluis 2018; Meghir, Palme, and Simeonova 2018; Clark and Royer 2013), nor can this explanation reasonably explain the sharp increase in the education-mortality gradient.

But there are other mechanisms by which life expectancy gaps may diverge. One would be lifestyle factors at the individual level, as Case and Deaton show in this paper. For example, the rising gap in marriage rates between people with a BA and those without a BA would be expected to increase the mortality gap given the beneficial health effects of marriage (Rendall and others 2011), but it's unlikely to explain the acceleration since 2010. Other potential factors include physical and social environments, policies, and social values (Woolf and Aron 2013). Still, one would expect that if local and state policies were key determinants of the rising educational gradient, as in Montez and others (2019, 2020), we would expect to see heterogeneity in the evolution of the education gradient across states, a hypothesis considered in the next section.

4. Centers for Disease Control and Prevention, National Center for Health Statistics, "Deaths by Week and State," https://www.cdc.gov/nchs/nvss/vsrr/COVID19/index.htm, accessed November 20, 2023.

Figure 1. Midlife Mortality Rates for Noncollege Graduates by Year Relative to the 1992 Baseline Mortality Rate for the Five States with the Largest Increase and Five States with the Greatest Decline



Source: Archived data from Couillard and others (2021).

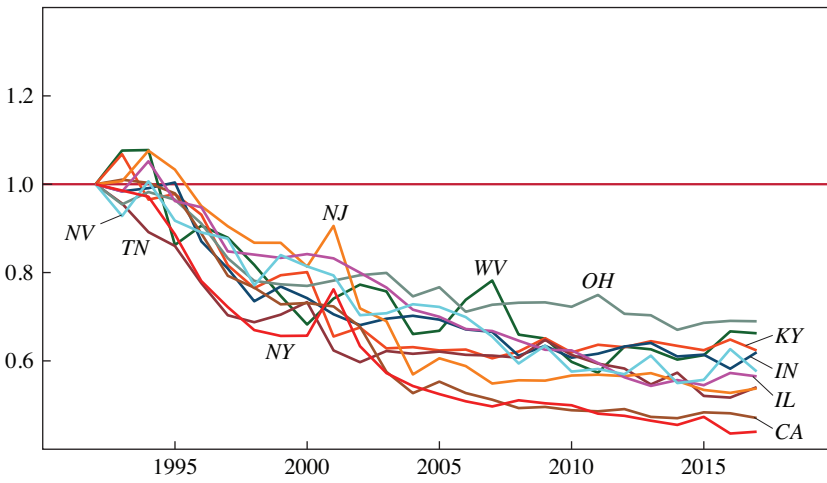
Note: The remaining thirty-four states in the sample would be inside the gap between the two groups of states.

STATE-LEVEL VARIATION IN COLLEGE AND NONCOLLEGE MIDLIFE MORTALITY

I use data from Couillard and others (2021) on midlife mortality by state, year, and education between 1992 and 2017 for forty-four states with complete data, and thus do not address the influence of COVID-19 on state-level mortality.[5] As well, the data focus only on midlife mortality (25–64) and so miss the evolution of mortality for older populations. To give a sense for the patterns across states, I consider in figure 1 the five states that experienced the greatest increase in mortality for noncollege graduates between 1992 (the reference year) and 2017—West Virginia, Kentucky, Indiana, Ohio, and Tennessee—and for the five states with the greatest relative decrease—New York, California, Nevada, New Jersey, and Illinois. While the fanning out is by design (the remaining thirty-four states are between these two groupings), it still indicates the divergence across states, with some experiencing more than 30 percent growth in midlife mortality (Kentucky and West

5. All calculations are based on the archival data supporting Couillard and others (2021): https://www.openicpsr.org/openicpsr/project/144041/version/V1/view.

Figure 2. Midlife Mortality Rates for College Graduates by Year Relative to the 1992 Baseline Mortality Rate for the Five States with the Largest Increase and Five States with the Greatest Decline



Source: Archived data from Couillard and others (2021).

Virginia), while for California and New York, the declines were 24 percent and 30 percent, respectively.[6]

Figure 2 shows the same trends in mortality for college graduates in each of the ten states considered above. While the top and bottom five states exhibit generally similar rankings, it is striking how closely the mortality patterns for college graduates track together; those in Tennessee—one of the five states with the greatest increases in noncollege-graduate mortality—experienced a decline in mortality for those with college educations equal to 46 percent, similar in magnitude to New Jersey (46 percent), Illinois (44 percent), and Nevada (42 percent). While a considerable degree of dispersion across states in mortality remains for college graduates (the standard deviation of log mortality in 2017 is similar for college and noncollege graduates), it is apparent from figure 2 that on average by state, people with a college degree experienced an expanded lifespan regardless of where they lived; the same could not be said for noncollege graduates.

It is increasingly clear that recessions were not the culprit for declining life expectancy, whether for health more generally (Ruhm 2000; Finkelstein

6. See also Montez and others (2019).

and others 2024) or as an explanation for the widening mortality gaps by education (Case and Deaton 2017). Indeed, Couillard and others (2021) found that even decades-long changes in regional income or unemployment (1993–2017) were uncorrelated with changes in contemporaneous log mortality rates. But the change in mortality is correlated with the initial level of state-level income in 1992 (a correlation coefficient of −0.58 for noncollege graduates and −0.54 for college graduates, both highly significant), and with the 1968 state-level income. This puzzling correlation is consistent with the work by Montez and others (2020), who have argued that the long-term effects of state-level policies such as tobacco taxes and smoking bans, minimum wages (and local minimum wage bans), gun control, civil rights, Medicaid, and environmental policies have led to widening longer-term increases in mortality dispersion across states. That many of these policies were enacted in the late twentieth century by higher-income states, and that they would have the greatest impact on noncollege graduates, is certainly consistent with the empirical patterns we observe.

There are two methodological difficulties in assessing how (and whether) state-level policies affect (or are just associated with) secular changes in mortality rates. The first is figuring out whether state policies are causal or instead reflect individual health preferences of the state residents. For example, smoking rates between 1992 and 2017 fell by more in New York than in Mississippi; this was likely affected by policies in New York designed to reduce smoking such as its $4.35 tax per pack in 2016, compared to those in Mississippi, with a 2016 tax of $0.66 (Couillard and others 2021). But it also likely reflects the preferences of New Yorkers both for less smoking and support for state legislation to reduce smoking (Besley and Case 2000). It's not clear that the package of New York policies would have had the same health effects had it been enacted in Mississippi.

Second, state policies are likely to affect health outcomes with (to paraphrase Milton Friedman) a long and variable lag. Teenage smoking restrictions and generous Medicaid benefits for children are unlikely to reduce mortality until many decades in the future; similarly, heavy drinking often takes many years to translate into premature death. (The exceptions are for opioid overdoses and suicides.) Combined with the potential endogeneity of state-level policies noted above and the very large number of state-level policies (well more than the number of states), this makes estimating the causal impact of state policies difficult. Still, Montez and others (2020) have pointed to the state-level private labor restrictions, tobacco taxes, environmental regulations, and civil rights legislation (among other factors) as those making the largest contributions to mortality reductions.

By the same token, we might expect the "long-term deterioration in opportunities for less educated Americans" (Case and Deaton 2020, 144), independent of state policies, to exhibit long and variable lags with respect to their impact on mortality rates. The long-term impact of stress arising from the loss of stable well-paying jobs, domestic instability, and the loss of community networks during the 1980s and 1990s is likely to contribute to the reversal of the previous growth in life expectancy, particularly for diseases such as cardiovascular disease or cancers (Geronimus and others 2019).

DISCUSSION This most recent study by Case and Deaton has documented an important and disturbing trend in the education gradient since the early 1990s, with an acceleration in the gap between college and noncollege graduates since 2010, especially during the COVID-19 pandemic. The authors have suggested several plausible mechanisms for why the gap has continued to grow. While parsing out individual effects is difficult, I agree with Case and Deaton that there were multiple causes for the rapidly widening gradient—a perfect storm of several correlated adverse factors, with the most recent being COVID-19.

Understanding why the education-mortality gradient continues to expand is important, especially in predicting whether it might stabilize or even reverse course after expanding for the past six decades. While we have a comprehensive list of suspects, untangling the influences of wages, labor force participation, factory closings, connections to the community, health care quality, health behaviors, local policies, and domestic living arrangements is difficult. But even a partial understanding of the state policy effects can contribute to lives saved in the future.

With the sharply declining mortality rate from COVID-19, it's likely that the jump in the education-mortality gradient arising from COVID-19 will become attenuated, with a disproportionate benefit to older people most at risk of COVID-19. That older people are affected disproportionately by COVID-19, while younger people are affected by deaths of despair, does lead to a broader point that the effectiveness of specific policies will be quite different for those at midlife (25–64) compared to those for older (65–84) people. For example, initiatives designed to reduce deaths of despair or to reduce future cancer and cardiovascular disease mortality, are more effective for younger populations by encouraging stable employment, domestic stability, and healthy behaviors. While health habits and stability may be important for 75-year-olds, the challenges for this group—at least among those who survived to age 75—is to manage chronic diseases to improve quality of life and longevity, a very different set of policy priorities. In sum, the study by Case and Deaton has made it crystal clear the extent and

magnitude of the problem facing the United States regarding the widening disparities in life expectancy by education, as well as providing a road map for what factors are likely contributors to the gap. Figuring out how to reverse this trend in the education gradient should be a major priority for the federal government and state governments, as it seems unlikely that the trend will reverse on its own.

REFERENCES FOR THE SKINNER COMMENT

Besley, Timothy, and Anne Case. 2000. "Unnatural Experiments? Estimating the Incidence of Endogenous Policies." *Economic Journal* 110, no. 467: 672–94.

Case, Anne, and Angus Deaton. 2017. "Mortality and Morbidity in the 21st Century." *Brookings Papers on Economic Activity*, Spring, 397–443.

Case, Anne, and Angus Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton, N.J.: Princeton University Press.

Case, Anne, and Angus Deaton. 2021. "Life Expectancy in Adulthood Is Falling for Those without a BA Degree, but as Educational Gaps Have Widened, Racial Gaps Have Narrowed." *Proceedings of the National Academy of Sciences* 118, no. 11: e2024777118.

Clark, Damon, and Heather Royer. 2013. "The Effect of Education on Adult Mortality and Health: Evidence from Britain." *American Economic Review* 103, no. 6: 2087–120.

Couillard, Benjamin K., Christopher L. Foote, Kavish Gandhi, Ellen Meara, and Jonathan Skinner. 2021. "Rising Geographic Disparities in US Mortality." *Journal of Economic Perspectives* 35, no. 4: 123–46.

Cutler, David M., and Adriana Lleras-Muney. 2010. "Understanding Differences in Health Behaviors by Education." *Journal of Health Economics* 29, no. 1: 1–28.

Finkelstein, Amy, Matthew J. Notowidigdo, Frank Schilbach, and Jonathan Zhang. 2024. "Lives vs. Livelihoods: The Impact of the Great Recession on Mortality and Welfare." Working Paper 2024-14. Chicago: University of Chicago, Becker Friedman Institute. https://bfi.uchicago.edu/working-paper/lives-vs-livelihoods-the-impact-of-the-great-recession-on-mortality-and-welfare/.

Foote, Christopher, Ellen Meara, Jonathan Skinner, and Luke Steward. 2024. "Geography and the Widening Educational Divide in U.S. Midlife Mortality." Working Paper, Dartmouth College.

Galama, Titus J., Adriana Lleras-Muney, and Hans van Kippersluis. 2018. "The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence." Working Paper 24225. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w24225.

Geronimus, Arline T., John Bound, Timothy A. Waidman, Javier M. Rodriguez, and Brenden Timpe, 2019. "Weathering, Drugs, and Whack-a-Mole: Fundamental and Proximate Causes of Widening Educational Inequity in U.S. Life Expectancy by Sex and Race, 1990–2015." *Journal of Health and Social Behavior* 60, no. 2: 222–39.

Kitagawa, Evelyn Mae, and Philip M. Hauser. 1968. "Education Differentials in Mortality by Cause of Death: United States, 1960." *Demography* 5, no. 1: 318–53.

Kitagawa, Evelyn Mae, and Philip M. Hauser. 1973. *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge, Mass.: Harvard University Press.

Leive, Adam A., and Christopher J. Ruhm. 2021. "Has Mortality Risen Disproportionately for the Least Educated?" *Journal of Health Economics* 79: 102494.

Meara, Ellen R., Seth Richards, and David M. Cutler. 2008. "The Gap Gets Bigger: Changes in Mortality and Life Expectancy, by Education, 1981–2000." *Health Affairs* 27, no. 2: 350–60.

Meghir, Costas, Mårten Palme, and Emilia Simeonova. 2018. "Education and Mortality: Evidence from a Social Experiment." *American Economic Journal: Applied Economics* 10, no. 2: 234–56.

Montez, Jennifer Karas, Jason Beckfield, Julene Kemp Cooney, Jacob M. Grumbach, Mark D. Hayward, Huseyin Zeyd Koytak, Steven H. Woolf, and Anna Zajacova. 2020. "US State Policies, Politics, and Life Expectancy." *Milbank Quarterly* 98, no. 3: 668–99.

Montez, Jennifer Karas, Anna Zajacova, Mark D. Hayward, Steven H. Woolf, Derek Chapman, and Jason Beckfield. 2019. "Educational Disparities in Adult Mortality across U.S. States: How Do They Differ, and Have They Changed since the Mid-1980s?" *Demography* 56, no. 2: 621–44.

National Center for Health Statistics (NCHS). 2023. "Provisional Data Shows U.S. Drug Overdose Deaths Top 100,000 in 2022." Blog post, May 18, Centers for Disease Control and Prevention. https://blogs.cdc.gov/nchs/2023/05/18/7365/.

Novosad, Paul, Charlie Rafkin, and Sam Asher. 2022. "Mortality Change among Less Educated Americans." *American Economic Journal: Applied Economics* 14, no. 4: 1–34.

Pappas, Gregory, Susan Queen, Wilbur Hadden, and Gail Fisher. 1993. "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986." *New England Journal of Medicine* 329, no. 2: 103–9.

Rendall, Michael S., Margaret M. Weden, Melissa M. Favreault, and Hilary Waldron. 2011. "The Protective Effect of Marriage for Survival: A Review and Update." *Demography* 48, no. 2: 481–506.

Ruhm, Christopher J. 2000. "Are Recessions Good for Your Health?" *Quarterly Journal of Economics* 115, no. 2: 617–50.

Schwandt, Hannes, Janet Currie, Till von Wachter, Jonathan Kowarski, Derek Chapman, and Steven H. Woolf. 2022. "Changes in the Relationship between Income and Life Expectancy before and during the COVID-19 Pandemic, California, 2015–2021." *Journal of the American Medical Association* 328, no. 4: 360–66.

Woolf, Steven H., and Laudan Aron, eds. 2013. *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. Washington: National Academies Press.

GENERAL DISCUSSION James Stock commented on the role that obesity trends might play in the authors' results, noting that the level of obesity is higher among lower-educated individuals. He observed that according to Centers for Disease Control and Prevention's Adult Obesity Prevalence Maps,[1] the overall adult obesity rates of the worst five states mentioned in Jonathan Skinner's discussant remark—West Virginia, Kentucky, Indiana, Tennessee, and Ohio—are all in the higher tiers, while the rates for Skinner's best five states—California, New Jersey, Illinois, Nevada, and New York—are in lower tiers. He suggested that the differential trends in cardiovascular disease might be related to these facts.

Anne Case acknowledged the crucial role of obesity in public health in America but disagreed with the notion that it would play a key role in the trends the authors identified. She pointed to the fact that even as obesity has risen for decades, deaths from cardiovascular disease continued to fall; since then, progress has largely flatlined across the English-speaking world, despite distinct obesity trends in different countries and states. Case suggested that, ultimately, there is something going on with the relationship between obesity and cardiovascular disease that experts do not yet understand, and given these often conflicting trends, obesity was an unlikely culprit to be driving changes in differential mortality, even as it remained a pressing public health challenge.

Robert Gordon followed up on Stock's comment, noting that beyond the dividing deaths between "deaths of despair" and other causes, we should also consider whether deaths are related to personal responsibility. He suggested that the obesity-related diabetes and heart disease are examples of the latter, and there is a distinction between lack of economic access to health care (e.g., due to lack of insurance) and geographic distance from health care in rural areas.

Martin Baily agreed with Gordon's point about personal responsibility. He also remarked that if selection is not a major concern, the authors' policy recommendations should have included encouraging college attendance. He further expressed surprise that the increased health coverage associated with Medicare expansion would not have moderated the effects they found during the COVID-19 pandemic.

Betsey Stevenson pointed out that, selection or not, the aggregate data suggest the United States is falling behind. She emphasized that policy

1. Centers for Disease Control and Prevention, "Adult Obesity Prevalance Maps," updated September 21, 2023, https://www.cdc.gov/obesity/data/prevalence-maps.html.

and education could influence people's ability to make informed choices about their health and well-being. Stevenson also remarked that notwithstanding comments about selection, education might play a causal role in the trends discussed in the paper. In particular, she pointed to the role of higher education in individuals' ability to interpret information and make healthy lifestyle choices.

Responding to the comments by Gordon and Baily, Case first noted that obesity is a poor example of personal responsibility and should instead be considered an addiction: some individuals "soothe the beast" through alcohol or drugs, while others do so through unhealthy relationships with food. She further argued that, while it is true that people choose their behaviors, we must ask why those kinds of choices are disproportionately made by people without a college degree or those who lack economic resources and access to physical or mental health treatment.

Benjamin Harris highlighted the role of labor force attachment, noting that work is an important source of social interaction and intellectual stimulation along with wages. He suggested that since labor force participation varies by place and by educational attainment, the divergence in labor force participation among both older and younger workers might play a role in the effects presented by the authors.

Alan Blinder followed up on the issue of selection raised by the discussants. He first remarked that getting a BA is perhaps a component of getting one's life together, and that there might be important differences in personality between those who complete a BA and those who don't. He continued by expressing doubt about whether, if the share of the population who complete a BA continues to rise, the economy would be able to provide good jobs utilizing the skills taught in those degrees to 50 percent or more of the population.

Justin Wolfers suggested that there might possibly be a simple mathematical calibration exercise that could verify whether selection into BA programs could plausibly explain a significant share of the authors' results.

Stan Veuger also commented on the issue of selection that, to the extent selection is a problem in the paper, encouraging college attendance may have limited impact as a policy solution. He observed that restricted housing supply in high-income cities and states could explain some of the patterns identified in the cross-sectional data if people without a BA who can afford to live in the richest cities and states are increasingly positively selected. In reference to Baily's comment on Medicare expansion, Veuger mentioned that some researchers have found insurance expansions might also offer more access to opioids, thereby exacerbating opioid addiction, which could

explain some of these unexpected effects, even though these studies may not be particularly rigorous.

Pinelopi Goldberg cast doubt on the idea that selection was a major driver of the authors' findings, noting that, as the authors previously documented, the overall life expectancy in the United States is declining on average and this indicates real effects that selection could not plausibly have driven. She also returned to Skinner's comments on place in his discussion and emphasized that the interaction between college education and place may be key to understanding the authors' results. In particular, relating this idea to the trade literature, the China trade shock may not have resulted in lower wages or income in affected areas but did result in lower employment, worse mental health, and other outcomes that might lead to deaths of despair. Thus, income is not the main driver, she argued. Finally, Goldberg concluded that if selection is a concern, selection in terms of residential location would be more pronounced as it relates to higher education. She reiterated that the interaction between the two might be important in understanding the widening gap in mortality.

Case responded to the comments on selection, emphasizing that selection could not plausibly explain many of their findings. In particular, Case highlighted the evidence that although the share of women with a BA did not increase between 1950 and 1965, the rate of deaths of despair nevertheless rose from one birth cohort to the next over this span of birth cohorts. The gap in deaths of despair continued to rise whether or not the share of people with BA was rising. She explained that these findings indicate large and policy-relevant effects that could not plausibly be driven by selection. She also referred to papers by Arline Geronimus and David Cutler, which suggest that selection is not significant enough to drive the lion's share of findings on differential mortality.[2]

Angus Deaton began by discussing the issue of causality and selection. He reflected on the growing focus on causal inference in economics, noting that the development of these tools had helped fill important blind spots in the field, but he also expressed his discomfort with the profession's recent obsession with causality. Deaton stated that while precisely identifying causal channels could be important for prescribing policy remedies, causality is

2. Ellen R. Meara, Seth Richards, and David M. Cutler, "The Gap Gets Bigger: Changes in Mortality and Life Expectancy by Education, 1981−2000," *Health Affairs* 27, no. 2 (2008): 350–60; Arline T. Geronimus, John Bound, Timothy A. Waidmann, Javier M. Rodriguez, and Brenden Timpe, "Weathering, Drugs, and Whack-a-Mole: Fundamental and Proximate Causes of Widening Educational Inequity in U.S. Life Expectancy by Sex and Race, 1990–2015," *Journal of Health and Social Behavior* 60, no. 2 (2019): 222–39.

not the only—or the best—metric to judge a finding, and that even trends driven by selection can be of crucial importance. He also argued that their paper's evidence on the widening mortality gaps within birth cohorts—even those whose education did not change later in life—suggests that selection could not plausibly explain a large share of their findings. Of course, denying the importance of selection does not mean that a college degree directly causes better health.

Elaine Buckberg remarked that the Affordable Care Act does not seem to have ameliorated the divergence in outcomes even for illnesses that are not related to behavior and that this raises questions about disparities in timeliness, quality, and quantity of care. Specifically, she pointed out that insurance coverage is not a binary variable, and that issues including marginal charges for care, wait time for appointments, and access to preventative screenings might play an important role in health outcomes even within the group of individuals with insurance coverage.

Hoyt Bleakley inquired about these results related to earlier work by Case and others on the emergence in childhood of health differences by parental education and income. He suggested that policies intended to close health disparities among adults might be less effective given prior research on the lasting impact of early life conditions on health.

In reference to the long-standing hunt by education and health economists to identify whether there is a causal effect of education on health outcomes, Deaton remarked that one of the most important contributions to the debate came not from economists at all, but from the sociological "fundamental causes theory" developed by Jo Phelan and Bruce Link.[3] In the theory, Phelan and Link describe how the power and status that come with wealth, income, and education will affect one's health if and only if there is an opportunity through which health can be affected. To illustrate the importance of this idea, Deaton described how until about 1750 in England, there was no income or education gradient for death rates—the rates among aristocrats and nobles were roughly the same as those among the general population because there simply were not methods to stop disease-related mortality regardless of education and income. It was only after 1750, when these mechanisms started to become available, that the rich and powerful were able to take advantage and began living longer.[4] Deaton contended

3. B. G. Link and J. Phelan, "Social Conditions as Fundamental Causes of Disease," *Journal of Health and Social Behavior*, special edition (1995): 80–94.

4. Angus S. Deaton, *The Great Escape: Health, Wealth, and the Origins of Inequality* (Princeton NJ: Princeton University Press, 2013), chapter 2.

that a similar finding could apply to the portions of their paper dealing with cancer mortality. Education still does not significantly affect mortality from brain cancer, as there are not sufficiently effective treatments for education and income to begin affecting health in this area, and it has only recently begun to affect mortality from breast cancer and some other cancers, because innovations in screening and treatment have provided pathways through which women can use their education and income to bring down mortality rates. Mortality rates from breast cancer, once higher for more educated women, are now lower.

Case noted that while health care is a crucial piece of the puzzle, the problems underlying rising mortality gaps will not all be solved in a doctor's office: the United Kingdom is also experiencing differential mortality trends despite its universal health coverage. Making progress on these issues, she commented, would require broader social and economic changes beginning in early life, including addressing the challenges facing children who have already lost their parents to drugs, alcohol, and suicide.

RICHARD BALDWIN
IMD Business School

REBECCA FREEMAN
Bank of England

ANGELOS THEODORAKOPOULOS
Aston Business School

Hidden Exposure: Measuring US Supply Chain Reliance

ABSTRACT Supply chain problems, previously relegated to specialized journals, now appear in G7 Leaders' Communiqués. Our paper looks at three core elements of the problems: measurement of the links that expose supply chains to disruptions, the nature of the shocks that cause the disruptions, and the criteria for policy to mitigate the impact of disruptions. Utilizing global input-output data, we show that the US exposure to foreign suppliers, and particularly to China, is "hidden" in the sense that it is much larger than what conventional trade data suggest. However, at the macro level, exposure remains relatively modest, given that over 80 percent of US industrial inputs are sourced domestically. We argue that many recent shocks to supply chains have been systemic rather than idiosyncratic. Moreover, systemic shocks are likely to arise from climate change, geoeconomic tensions, and digital disruptions. Our principal conclusion is that the concerns regarding supply chain disruptions, and policies to address them, should focus on individual products rather than the whole manufacturing sector.

Conflict of Interest Disclosure: The authors did not receive financial support from any firm or person for this paper or from any firm or person with a financial or political interest in this paper. The authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper. The Bank of England, Rebecca Freeman's employer, had the right to review this work for sensitivity screening prior to publication. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees.

Brookings Papers on Economic Activity, Fall 2023: 79–134

When Harold Macmillan—the UK prime minister in the turbulent post-WWII years—was asked: "What is the greatest challenge you face?" his alleged reply was: "Events, my dear boy, events." Events, termed "shocks" by economists, have reemerged as formidable challenges for global leadership, with supply chain disruptions being top of mind. At their May 2023 summit, for example, G7 leaders stated that "supporting resilient and sustainable value chains remains our priority" (European Council 2023, 1). It was not always like this.

Constructed in a time of stability and hope, today's globe-spanning supply chains propelled efficiency and progress as they became the arteries of the US economy. US administrations supported the internationalization of supply chains with the entry into force of deep trade agreements, like the North American Free Trade Agreement on January 1, 1994, and the establishment of the World Trade Organization on January 1, 1995. At the time, international supply chains were viewed as enhancers of productivity and boosters of prosperity (CEA 2016).

But supply chains are behaving differently in the face of what Mervyn King and John Kay term "radical uncertainty" in their 2020 book of the same name. Today, reverberations of supply chain disruptions echo loud and long, influencing everything from laptop availability and headline inflation to national security and shortages of medicine that affect millions. Empirical studies of these effects are just emerging (Goldberg and Reed 2023; Boehm, Flaaen, and Pandalai-Nayar 2019; Carvalho and others 2021; Bonadio and others 2021). Most of the economic literature on global supply chains (GSCs) study factors that foster them (Grossman and Rossi-Hansberg 2008; Antràs 2020; Alfaro and Chor 2023) or investigate broader scale trends in the landscape of GSCs (World Bank 2020). Economic research on supply chain disruptions is appearing on the theory side (Grossman, Helpman, and Lhuillier 2021; Carvalho and Tahbaz-Salehi 2019; Elliott and Golub 2022; Elliott, Golub, and Leduc 2022; Baqaee and Rubbo 2023) and on the empirical side (Schwellnus, Haramboure, and Samek 2023a; Imbs and Pauwels 2022).

As these are early days for the economics of supply chain disruptions, there is no consensus on how to organize thinking about the related issues. We propose that the phrase "supply chain disruptions" inherently directs us toward a three-pillar organizing framework: the links that constitute GSCs, the shocks that disrupt them, and policies that mitigate or avoid the disruptions. Our paper is organized around these three pillars.

The rest of the paper comes in five sections. Section I looks at how we can measure the links. Section II shows our empirical findings on the

exposure of US manufacturing sectors to domestic and foreign supply chain links, with a special focus on China (the largest foreign supplying nation). Sections III and IV present, respectively, frameworks for thinking about shocks and policy. Our concluding remarks are in section V.

I. The Links: On the Measurement of Supply Chain Exposure

In US manufacturing companies, supply chain risk managers have long recognized the importance of knowing their suppliers (Gurtu and Johny 2021). However, the advent of supply chain disruptions on a grand scale, spanning multiple sectors and nations, has elevated this issue from a firm-level concern to a nation-level concern. Identifying where things are actually made, however, is not as easy as it might appear.

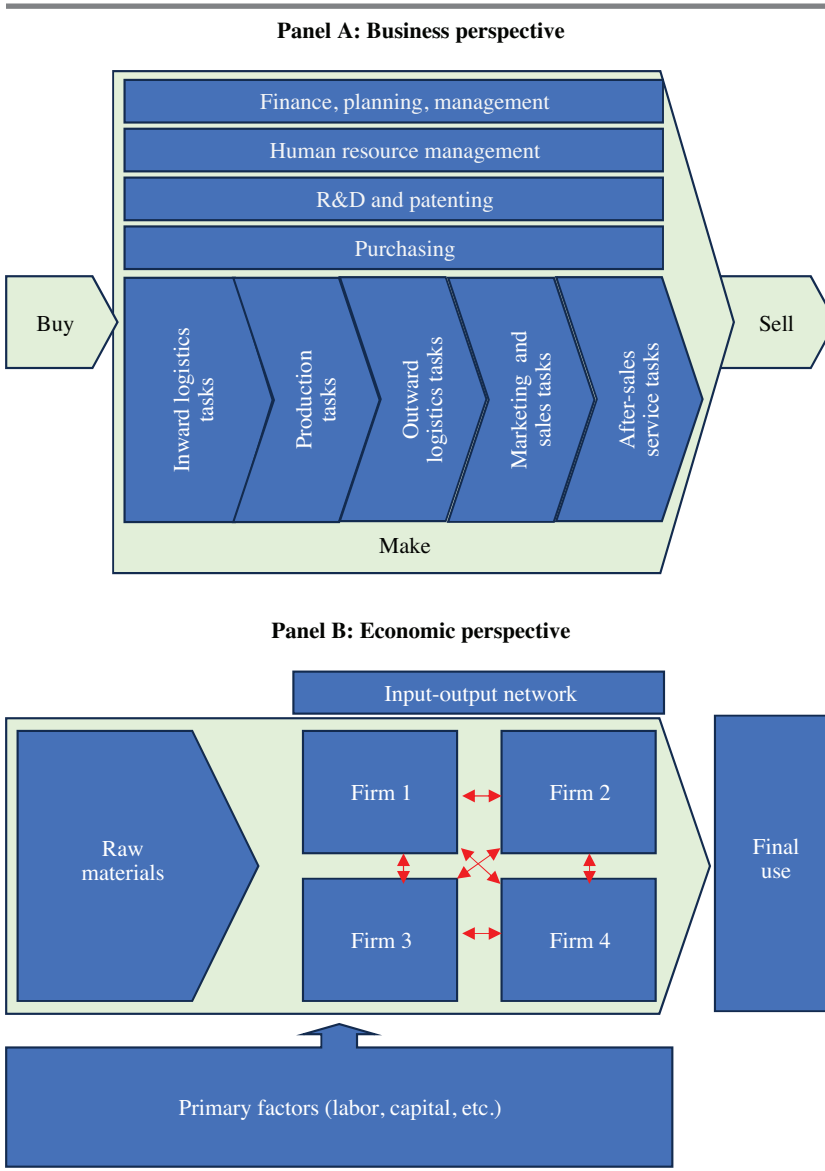
I.A. You Can't Fix What You Can't See: Two Ways of Looking at Supply Chains

A cornerstone contribution of our paper lies in the identification of the true origin of the manufactured inputs bought by US manufacturing sectors. We are not the first to tackle the problem. Many studies have taken what could be called the business value chain approach to trace out a firm's supply chain. Our paper presents measures of supply chain exposure that rely on a very different approach.

BUSINESS VALUE CHAIN APPROACH VERSUS ECONOMIC APPROACH Much of the excellent, detailed work on supply chain dependencies has used the business, or value chain approach. The Biden administration, for example, has set up a series of initiatives to map industrial supply chains (White House 2022) with an eye to revealing where potential weak points may lie. These initiatives take a business-focused approach inspired by Michael Porter (Porter 1985). At its core, this is based on a straightforward view that firms buy things to make the goods that they sell. The direct suppliers are called tier 1 suppliers, their suppliers are called tier 2 suppliers, and so on. This approach establishes a sequence, or chain, of supplying firms, which is why the literature uses the phrase "supply chain," or "value chain" when speaking about the network of suppliers (figure 1, panel A). This is quite different from the economic approach, as panel B of figure 1 illustrates.

Economists tend to take a bird's-eye view. The buy-make-sell logic of Porter's value chain is recursive, establishing an input-output network of firms selling to firms and eventually to final customers (figure 1, panel B). This shows that what looks like a chain of suppliers for a single firm is,

Figure 1. Supply Chain Perspectives: Business Value Chain versus Economic Input-Output Table



Source: Authors' elaboration of the Porter (1985) value chain (panel A) and a schematic view of a firm-level input-output table (panel B).

in fact, part of a matrix from the economy-wide perspective. In addition, the economic viewpoint introduces a distinction between primary inputs like labor and capital, intermediate inputs such as parts and components, and final goods.

One way to conceptualize the differences between the two methods is to consider the analogy to the differences between family trees that serve as a parallel to the business approach, and broader genealogical approaches such as social network analysis, which are akin to the economic approach. In the context of family trees, parents can be viewed as tier 1 suppliers, grandparents as tier 2 suppliers, and so forth. Although family trees provide a valuable means of identifying key familial connections, they are insufficient for grasping the complexities of broader communities. For a more comprehensive understanding, social network analysis is essential.

The business and economic approaches each have their advantages.[1] The business view allows much greater attention to detail as panel A in figure 1 makes clear. By focusing on a single firm, an analyst can delve deep into issues such as logistics, inventory control, and risk management strategies as well as the required administrative tasks ranging from financial planning to purchasing policies (horizontal bars in panel A). Additionally, they can concentrate on corporate relations, partnerships, contracting, and product portfolios. If the ultimate policy goal is to avoid disruption of production of a particular good, say, semiconductors, the business approach is the one to take. It is like following a river from its mouth back to the source of all its tributaries. This approach, however, would not have picked up the shock to US car production in 2020 that came when the demand for semiconductors boomed from other sectors, like work-from-home equipment. For that, an economy-wide perspective is necessary.

THE CORE DIFFICULTY AND THE TWO SOLUTIONS The two approaches, while quite different, face a common core difficulty: the massive complexity of modern supply chains. The business approach and the economic approach take very different paths in addressing this core difficulty. An illustration using the auto industry clarifies the two solutions, each of which involves ignoring certain aspects of the complexity.

The business approach example comes from Lund and others (2020). This study found that General Motors (GM) had 856 tier 1 suppliers, but these 856 suppliers had suppliers themselves, the so-called tier 2

1. At a conceptual level, the two perspectives can also be combined. For instance, drawing upon Feenstra (2009), Fort (2023) presents a framework for firm decisions to engage internationally and outsource tasks.

suppliers, as did the tier 2 suppliers, and so on. The research estimated that GM had a staggering 18,000 suppliers in tier 2 and below. Given that each of these 18,000 suppliers had its own roster of suppliers, an exhaustive cataloging of GM's suppliers would create a sequence that reaches what Buzz Lightyear would call "infinity and beyond."

The business approach keeps the complexity manageable by drawing the line at the number of tiers investigated. The economic approach takes a very different method to the Buzz Lightyear problem, a very different approach to the suppliers of the suppliers, and embraces a very different type of simplifying assumption. The key is an analytic tool called input-output (IO) analysis, which works at the level of sectors rather than firms. The payoff from this simplification—aggregating all firms into sectors—is that IO analysis can deal fully with the suppliers-of-suppliers challenge. We illustrate this with the US car industry.

WHERE US-MADE CARS ARE MADE: ECONOMIC APPROACH In the economic approach, there are three levels of answers to the question, "Where are Ford cars actually made?" The first level is the easiest: one obvious answer is Dearborn, Michigan. When a Ford rolls off the assembly line in Dearborn, Michigan, we can say that the car was made in Dearborn. This is true, but it is not the whole truth. The second level admits that the Dearborn plant buys car parts from other firms. Many of those parts are not made in Michigan, and many are not made in the United States. Some are made in Canada, so we can say that some of the Dearborn-made cars were actually made in Canada. This is also true, but still not the whole truth. The third level digs into the fact that all the parts makers also buy parts—some of which are not made locally. Canadian car-part makers, for example, may source parts from Germany.

The problem is that the third level involves the same sort of Buzz Lightyear never-ending sequence encountered by the business approach. Parts makers buy parts from other parts makers that buy parts from other parts makers, and so on without end. IO analysis tackles the infinite recursion problem with matrix algebra.

I.B. Measuring Supply Chain Exposure with IO Tables

IO analysis, developed by Nobel laureate Wassily Leontief in the 1950s, shows how production in each sector relies on inputs from all sectors.[2] The

2. See Baldwin, Freeman, and Theodorakopoulos (2022) for a fuller discussion of IO analysis.

international version we use in this paper, the 2021 release of Inter-Country Input Output (ICIO) tables by the Organisation for Economic Co-operation and Development (OECD), tracks all sectors in the sixty-five countries in the data along with a rest-of-world aggregate.[3] A limitation of IO analysis is that it is conducted at the level of sectors and nations, so we cannot disaggregate down to the product or firm level. Moreover, because the data sets require detailed mapping and harmonization of data from national, regional, and international sources for different countries and across many time periods, IO data typically exhibit a larger lag in availability than, say, standard data on direct trade flows. For instance, the ICIO tables are available from 1995 to 2018.[4] There are efforts underway to use "nowcasting" methods to project IO calculations for the most recent years (even without complete data), but these are experimental at this stage (Mourougane and others 2023). In our view, the starting date is not a major issue since the expansion of offshoring and the "new" globalization began in earnest in the 1990s (Baldwin 2016). The end date is also less constraining than one might initially think because the COVID-19 pandemic caused significant disruptions to the global manufacturing and distribution networks, which are now stabilizing.

The heart of our analysis is the IO table and the distinction between goods that are used as intermediate inputs into the production of other goods (business-to-business, or B2B sales) and final goods sold to end users (business-to-customer, or B2C sales). The sum of a sector's sales of intermediate and final goods is called gross production, to distinguish it from net production, which corresponds to the output of final goods. Roughly speaking, gross production is the sector's total business turnover, or value of total sales. To avoid confusion, it is important to keep in mind that a sector both buys and sells intermediates. In this paper, we focus on supply-side exposure and note that a single sector's supply chain dependency turns on its purchases of intermediates, not its sales of intermediates. We could also look at the dependency on the selling side and work out a sector's dependence on supply chains for its sales.[5]

3. OECD, "OECD Inter-Country Input-Output Database," https://www.oecd.org/sti/ind/inter-country-input-output-tables.htm.

4. These tables form the basis for the OECD's Trade in Value Added (TiVA) database. Note that a new version of the OECD ICIO data, which comprises additional countries and two additional years, was released after the time that the analysis for this paper was conducted.

5. See Baldwin, Freeman, and Theodorakopoulos (2022) for discussion and calculations.

The IO table also shows the inputs that each sector in each country buys from every other sector in every other country. As such, the IO table has as many columns as rows, with each representing a sector in a particular country. The numbers in the table's cells represent the direct, or "face value," purchases by the column sector of inputs from the row sector. For example, the column in the IO table corresponding to the US Vehicles sector lists all the sector's purchases from all other sectors in every country. Using the second-level logic, the US Vehicles sector's purchases of inputs from other US sectors would be considered as made in the United States.

As it turns out, we can use IO analysis to solve the Buzz Lightyear, infinite sequence problem. With a series of simple yet unenlightening calculations, we can transform the IO table into the so-called Leontief matrix (see online appendix II for a more precise explanation). The elements of the Leontief matrix provide the third-level answer, in other words, the full links between all sectors and all nations, fully accounting for the fact that suppliers themselves have suppliers. To give it a name, we call the full accounting links "look-through" exposure.

FACE VALUE VERSUS LOOK-THROUGH EXPOSURE A critical feature of the economic approach is the distinction it makes between the face value exposure of a supply chain and its look-through exposure. Face value exposure measures look at the proximate origin of intermediate inputs. This corresponds to the second-level answer mentioned above that takes the origin of purchased intermediates at face value. For example, if an automaker in the United States buys a component from Canada, the face value exposure of the component is only to Canada. By contrast, the look-through exposure takes account of the fact that the Canadian producer of the component surely purchased inputs from other nations. In other words, the look-through exposure pierces the veil of the supplier network of suppliers supplying suppliers to identify the comprehensive link between a purchasing sector in one nation and every supplying sector in every nation.

As we shall see below, there is a substantial difference between supply chain exposure to some economies—especially China—when the exposure is measured on a look-through basis versus a face value basis.

LIMITATIONS OF INPUT-OUTPUT ANALYSIS A significant limitation of IO analysis is its omission of elasticities and lack of consideration for substitutability. For instance, the US textile industry heavily relies on imported inputs, many of which either originate in China or are produced using materials from China. At first glance, one might infer that this US sector is susceptible to disruptions. However, it is important to note that numerous countries export textiles and apparel. Consequently, any supply chain

disruptions can often be quickly mitigated by switching to alternative suppliers. Additionally, the relatively straightforward nature of these products makes switching suppliers in this sector easier than with more complex components, such as transmissions for trucks.[6]

Recent work, for example, by Moll, Schularick, and Zachmann (2023), also highlights how substitutability and agility can help prevent full-blown supply chain crises. Drawing lessons from Germany, they point to the role that the European market played in mitigating gas shortages after Russia curtailed its supply, beginning in 2021, thus preventing full-blown supply chain shutdowns. While there is evidence that elasticities of substitution at the micro level are known to be smaller than at the macro level (Houthakker 1955; Jones 2005; Oberfield and Raval 2021), readily available elasticities—especially for intermediates—would allow us to study the quantitative links between GSC disruptions and economic outcomes in a more meaningful manner. Goldberg and Reed (2023) make the related point that one would need information on all the elasticities of substitution at a highly disaggregated level to properly assess a product market's ability to withstand a given shock.

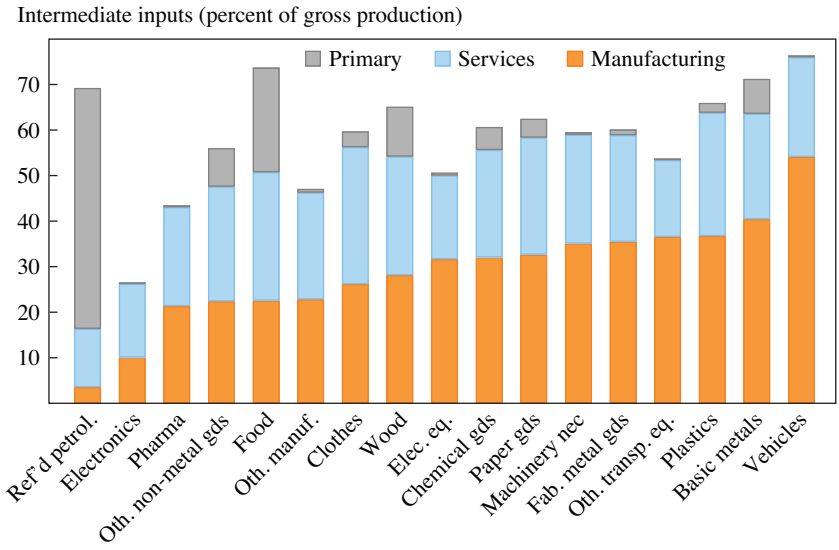
Furthermore, as mentioned above, an additional limitation of IO analysis is that it is conducted at the level of sectors and nations. Given the stringent requirements to construct IO tables, the data do not currently permit disaggregation down to the firm (or even detailed product) level, especially when multiple countries are included. As such, the economic repercussions of supply chain exposure, as it can be measured with the available IO data, may differ depending on the firm-level configuration of the supply chain (Baqaee and Farhi 2019; Elliott and Golub 2022).

II. The Links: Facts on United States and Comparator Nations

Some sectors, such as the auto sector, are inherently intensive in their use of purchased inputs and thus intrinsically more vulnerable to supply chain disruptions. To set the baseline for our study of foreign exposure, we look at the exposure of US manufacturing sectors to inputs from all sources, domestic and foreign, using the face value concept.

6. Antràs (2020) and Antràs and Chor (2022) note the sticky nature of supply chains and B2B relationships, which could in principle make it difficult to switch suppliers readily. However, some of the "stickiness" referred to is precisely generated by the lack of alternative suppliers, which is not the case for all sectors, as well as the need for complex, highly specialized parts and components, which are either not required or can more easily be replaced imperfectly for the production of some final goods.

Figure 2. Supply Chain Exposure of US Manufacturing Sectors by Type of Input, 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: The numbers shown represent the value of intermediates sourced by each US sector as a share of its total gross production.

II.A. Supply Chain Exposure of US Manufacturing Sectors

In the data upon which we draw—the 2021 release of the OECD ICIO tables[7]—we measure US purchased inputs in dollars and standardize each sector's input purchases by its gross production to allow comparisons across sectors and over time. Figure 2 presents the data for the year 2018, the most recent year in the data set. The chart displays stacked columns for each of the seventeen US manufacturing sectors identified in the database (see online appendix IV for a description of the products associated with various sectors).[8] The total height of each column reflects the importance

7. OECD, "OECD Inter-Country Input-Output Database," https://www.oecd.org/sti/ind/inter-country-input-output-tables.htm.

8. For convenience, we use shortened sector names as follows: Food products, beverages and tobacco = Food; Textiles, textile products, leather and footwear = Clothes; Wood and products of wood and cork = Wood; Paper products and printing = Paper gds; Coke and refined petroleum products = Ref'd petrol.; Chemical and chemical products = Chemical gds; Pharmaceuticals, medicinal chemical and botanical products = Pharma; Rubber and plastics products = Plastics; Other non-metallic mineral products = Oth. non-metal gds; Basic metals = Basic metals; Fabricated metal products = Fab. metal gds; Computer, electronic and optical equipment = Electronics; Electrical equipment = Elec. eq.; Machinery

of the sector's spending on intermediate inputs, counting inputs from all nations, including the United States itself. The bars within the columns indicate the broad source sectors of the intermediates. For clarity, we use the classic three-way classification of inputs, namely those coming from primary sectors (agriculture, mining, and utilities), services sectors, and manufacturing sectors. The sectors have been arranged in ascending order of their utilization of manufactured intermediate inputs.

For example, intermediate inputs amount to about 75 percent of the gross output of the Vehicles sector. How should we think about this 75 percent figure? Gross output in our data is measured in dollars and is defined as the sum of all costs, viewing profit as a payment to a factor of production and thus a cost. The costs comprise payments to factors of production (labor, capital, etc.) and purchased inputs (i.e., intermediate goods). The 75 percent figure means that, for the Vehicles sector, intermediate purchases make up three-quarters of all the costs. That is a very large number, and it means that the US Vehicles sector is highly exposed to supply chain issues—both domestic and foreign.

Note that intermediate inputs account for over half the costs in fourteen of the seventeen sectors. Even the sector with the lowest dependency, Electronics, has about 25 percent of its production cost linked to suppliers. Moreover, this 25 percent figure has to be handled with care since it is only for US manufacturers. At the global level, the Electronics sector is one of the most intensive users of intermediate goods, but the United States makes only a narrow range of the goods. Thus, the sector's low dependence shown in figure 2 arises from selection issues, not a ground-level reality of production processes. A similar point applies to the US pharmaceutical industry. In this sector, goods produced in the United States rely on intellectual property, which in the IO table and figure 2 registers as a service sector input.

Much of the recent discussion turns on manufactured inputs purchased by the manufacturing sector, so we zoom in on industrial inputs. Examining each sector's reliance on manufactured inputs, it is useful to divide the seventeen sectors into those with above- and below-median dependence on manufactured inputs. Notably, the sectors with above-median supply chain exposure include Electrical Equipment, Chemical Goods, Paper Goods,

---

and equipment, nec = Machinery nec; Motor vehicles, trailers and semi-trailers = Vehicles; Other transport equipment = Oth. transp. eq.; Manufacturing nec; Repair and installation of machinery and equipment = Oth. manuf.

Machinery nec (not elsewhere covered), Fabricated Metal Goods, Other Transport Equipment, Plastics, Basic Metals, and Vehicles. At the other end, the sectors that display below-median dependence are Refined Petroleum, Electronics, Pharmaceuticals, Other Non-Metal Goods (glass and ceramic products, construction materials, etc.), and Food.

Intermediate inputs originating from services sectors are also of interest. While usually seen as less vulnerable to shocks than industrial inputs, specific services such as cloud services, might pose significant risks for certain manufacturers. We have recently argued that trade in intermediate services is likely to dominate future trade (Baldwin, Freeman, and Theodorakopoulos 2023), but as of yet, they are not very important in the United States, so we set them aside for the rest of this paper.

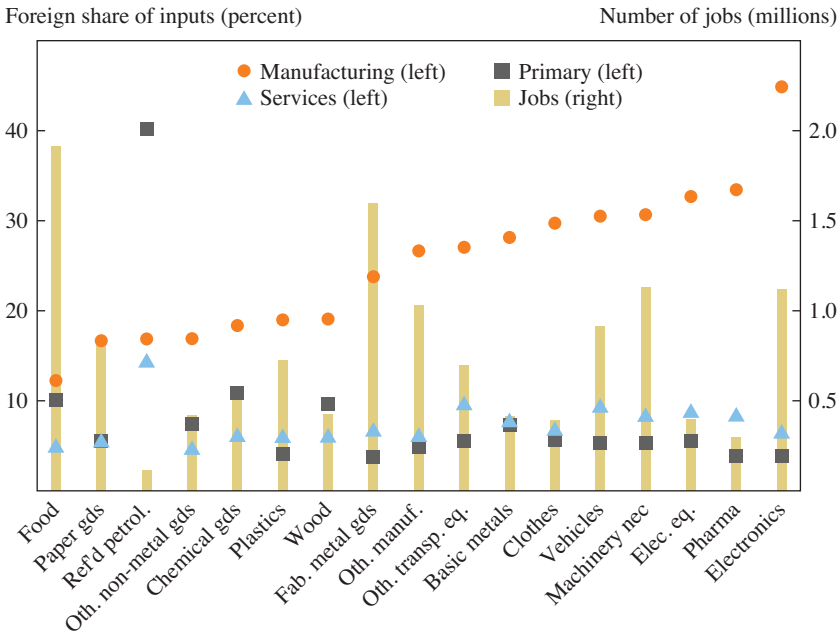
Regarding primary inputs, the observed patterns align with expectations. Primary inputs play a substantial role in only a handful of manufacturing sectors, including Refined Petroleum (53 percent), Food (23 percent), and Wood (11 percent). Surprisingly, the Basic Metals sector, known for producing items like steel girders, aluminum sheets, and copper wire, exhibits a relatively smaller share of inputs from primary sectors (8 percent). This can be attributed to the fact that, in the United States, much of the bulk production of basic metal goods relies on processing scrap metal rather than mining. As the collection and wholesaling of scrap metal are considered services, the US Basic Metals production depends less on primary sector inputs than one might assume.

II.B. Foreign Supply Chain Exposure by Sector at Face Value

Here we shift the focus to foreign sources of intermediate inputs—continuing to use the face value concept. Before looking at the facts, it is important to put the notion of foreign exposure into context to dispel the idea that foreign suppliers are somehow innately riskier than domestic suppliers. The point is that the riskiest thing to do with supply chains is to put all your eggs in one basket, even when the basket is at home (Miroudot 2020b; Baldwin and Freeman 2020a). Diversification of suppliers at home and abroad can be a useful buffer against shocks. During the pandemic, for example, having access to foreign suppliers was critical to reduce the disruption caused by domestic demand shocks in medical products (Evenett 2021). In short, the simplistic view that domestic suppliers are safe and foreign suppliers are risky is just that—simplistic.

Turning to the numbers, figure 3 unpacks the facts from figure 2 by displaying the foreign sourcing in each of its stacked bars. For example, manufacturing inputs for the rightmost column in figure 2 shows the share

Figure 3. Foreign Share of Intermediate Inputs by Type and Number of Jobs, United States, 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables and OECD 2021 Trade in Employment (TiM) database.

Note: This figure shows the sector's foreign purchased inputs as a share of its total inputs (domestic and foreign) by type of input (left axis) and number of US jobs (right axis).

of industrial inputs in the cost of production in the Vehicles sector. The Vehicles point for manufactured inputs in figure 3 indicates that 31 percent of these inputs are sourced from abroad. The domestic share is naturally the balance between the foreign share and 100 percent.

The first thing to note is that the focus of the recent public debate on industrial inputs—as opposed to, for example, primary inputs—seems justified. Apart from Refined Petroleum, foreign exposure to inputs in the primary and tertiary sectors is rather limited; the foreign share for these types of goods is generally less than 10 percent. As such, the rest of this paper focuses exclusively on the role of manufactured inputs in supply chains. A second key fact that emerges from figure 3 is the similarity of the foreign exposure shares when it comes to manufactured inputs. Apart from Electronics, which has a very high foreign share (45 percent), and Food, which has a very low foreign share (12 percent), the US manufacturing

sectors source between 16 percent and 33 percent of their manufactured inputs from abroad, with the median imported share being 27 percent. The foreign share is above the median for Other Transport Equipment, Basic Metals, Clothes, Vehicles, Machinery nec, Electrical Equipment, Pharmaceuticals, and Electronics. Nine of the seventeen sectors have foreign shares over a quarter.

The fact that the median foreign share is 27 percent means that most US sectors source the majority of their inputs from domestic suppliers. This is to be expected. As is true of all mega-economies, the United States is quite self-sufficient in industrial inputs (Baldwin and Freeman 2022). The explanation is straightforward. Empirical studies show that trade flows are very sensitive to distance; the rough rule of thumb is that bilateral trade flows fall by half when the distance between countries doubles (Head and Mayer 2014). Research also shows that the anti-trade effect—or to put it differently, the localization effect—of distance is even higher for intermediate goods (Miroudot, Lanz, and Ragoussis 2009; Conconi, Magerman, and Plaku 2020). The distance effect is countered by a size effect whereby countries trade more with big economies. It is natural, then, that the United States trades mostly with itself. It is, after all, a very large economy that is far from most nations, especially other large nations. Canada and Mexico are exceptions. Online appendix figure A1 shows this self-reliance in numbers. For the average US manufacturing sector, about 80 percent of all intermediates are sourced domestically. Thus, most of the United States' supply chain exposure is to itself.

When thinking about a sector's exposure to foreign suppliers and the implications that such exposure might have for the economy, a second set of important facts is the sector's size. Size, however, can be defined in many ways. Figure 3 shows the sectors' sizes as measured by jobs. The largest sector is Food, with almost 2 million employees in 2018. Fabricated Metal Goods is the second largest, with roughly 1.6 million jobs. Three other sectors employ more than a million people (Electronics, Other Manufacturing, and Machinery nec), but the rest of the sectors are comparatively small. Refined Petroleum, Pharmaceuticals, Clothes, Electrical Equipment, Basic Metals, Other Non-Metal Goods, and Wood all employ less than half a million workers.

II.C. Hidden Exposure: Look-Through versus Face Value Exposure

The next step is to look at exposure by sector and source nation, switching to the look-through basis to get the complete exposure of sectors to particular foreign suppliers. Our data set has sixty-five economies, but to

concentrate on the most important, we show the figures for only the top fifteen suppliers, which account for the lion's share of imported intermediates.

Figure 4 presents figures for the value of industrial inputs on a look-through basis, with the values standardized by the value of each sector's total purchases of manufactured intermediates from all sources—domestic and foreign. The supplying economies are listed in descending order of importance as a source, as measured by the simple average of the corresponding country's share in each of the seventeen manufacturing sectors (see rightmost column). To interpret the figures, note that, for example, the 5.1 percent in the Vehicles column for the China row indicates that China is the source of 5.1 percent of all manufactured inputs used by the US Vehicles sector on a look-through basis.

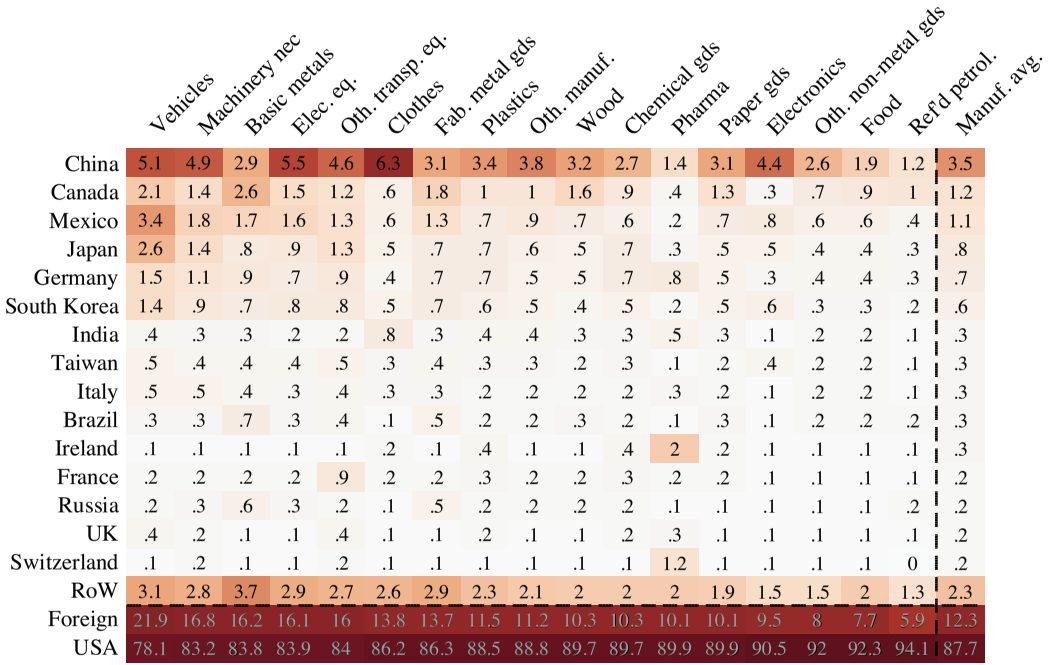
China's role as the dominant foreign supplier of industrial inputs to US manufacturing sectors is clear. Looking at the simple average across the seventeen sectors (rightmost column) shows a figure of 3.5 percent for China—close to three times larger than the average for the next closest supplier, Canada. Indeed, China's average share is more than the sum of the three next most important suppliers combined. In seven of the seventeen sectors, including Electrical Equipment, Plastics, and Fabricated Metal Goods, China is a more important supplier than the next four suppliers combined. In four of those sectors, China's share exceeds that of the next five most important suppliers. In two of these sectors, Clothes and Electronics, China's share exceeds that of the other top ten suppliers. This reflects the fact that China is also the top supplier for most of the United States' other top suppliers (Baldwin, Freeman, and Theodorakopoulos 2022).

Canada is particularly important as a supplier in Vehicles, Basic Metals, and Fabricated Metal Goods. Mexico is the third most important supplier followed by Japan, Germany, and South Korea. Once we get beyond the top six supplying economies, the only large suppliers are Ireland and Switzerland in the Pharmaceuticals sector (each accounting for more than 1 percent of inputs).

Our look-through measure also tells us that it is not just the United States that is heavily dependent on China for industrial supplies. Baldwin, Freeman, and Theodorakopoulos (2022), for example, show that in addition to the United States, all other major manufacturing nations source at least 2 percent of their total industrial intermediates from China.[9]

9. The nations included are Canada, Mexico, Germany, the United Kingdom, France, Italy, Japan, South Korea, and India.

Figure 4. Look-Through Exposure to Foreign Manufactured Intermediates (Percent), 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: Look-through exposure is the foreign production exposure: import side (FPEM) indicator as described in Baldwin, Freeman, and Theodorakopoulos (2022) and is computed as the share of manufactured inputs sourced by a given US sector from a given country on a look-through basis in total manufactured intermediates across all sources (foreign and domestic) on a look-through basis. "RoW" stands for the rest of the world. "Foreign" is the sum of all foreign sources.

To highlight the hidden exposure in US supply chains, figure 5 presents the percentage point difference between look-through exposure in figure 4, and the equivalent numbers for face value exposure.[10] The biggest differences are in sectors that are marked by extensive global supply chains. In such sectors, the hidden value gets added at many stages of the globalized production process. The differences are particularly marked in Vehicles, Machinery nec, Electrical Equipment, and Clothes. As far as source nations are concerned, the biggest hidden value is for nations that are important producers of intermediate goods and extensively involved in global supply chains. This includes the major manufacturing nations, which are (apart from the United States) China, Germany, and Japan.

The hidden exposure is very large. For example, the Vehicles sector's exposure to Chinese industrial inputs is about four times higher than indicated by the face value measure. In fact, the Chinese look-through exposure is more than four times the face value exposure in eight of the seventeen sectors. The percentage point differences are, on average, still quite high for Canada, Mexico, Japan, Germany, and South Korea, as the rightmost column shows. The only other big hidden exposure numbers are for Ireland and Switzerland in Pharmaceuticals.

II.D. Hidden Exposure Take 2: Rapid Concentration of Foreign Sourcing

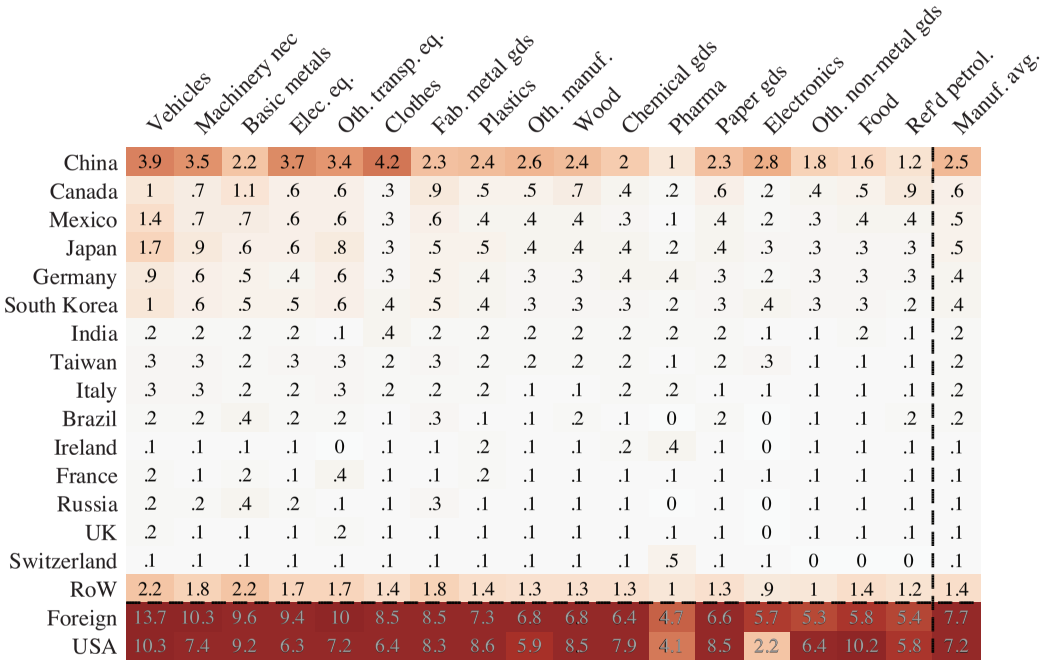
The "hidden" in hidden exposure in the previous section referred to the sourcing of intermediate inputs that was masked behind the Buzz Lightyear spiral of inputs used to make inputs. Here we turn the spotlight on another form of hidden exposure, namely the rapid geographic concentration of supply chain exposure.[11] It could be considered as hidden in the sense that it may have been underappreciated since it happened so fast.

CONCENTRATED SOURCING FROM CHINA The manufacturing of intermediates has rapidly become geographically concentrated in China. China's ascent as the world's top manufacturer is well documented (World Bank 2020). Less well-known is the fact that its production of intermediate manufactured goods has advanced even more rapidly than its production of final goods. Simply put, China has become what might be called the "OPEC of industrial inputs" (Baldwin 2022, par. 15). This concentration matters since supply chains fundamentally revolve around intermediate goods.

10. See online appendix figure A1 for the face value equivalent to figure 4.

11. In our analysis, we focus on concentration at the country level. However, it is worth noting that concentration can also exist within a given country. Data on the latter are typically not readily available at large scale.

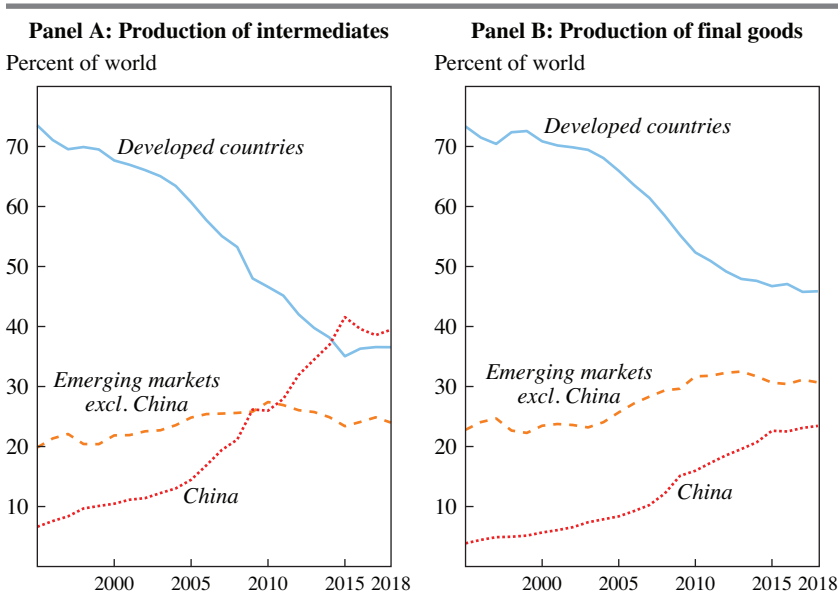
Figure 5. Hidden Exposure of US Sectors to Foreign Manufactured Intermediates (Percentage Point Difference between Look-Through and Face Value Exposure), 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: This figure presents the percentage point difference between the look-through exposure (figure 4) and face value exposure (online appendix figure A1). "RoW" stands for the rest of the world. "Foreign" is the sum of all foreign sources.

Figure 6. World Production of Intermediate and Final Manufactured Goods, 1995–2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: "Developed countries" include the European Union (EU), European Free Trade Association (EFTA) nations, the United Kingdom, the United States, Canada, Japan, Australia, and New Zealand. "Emerging markets excl. China" includes all other nations (including the rest of world aggregate) except China.

As figure 6 (panel A) shows, as recently as 1995, more than 70 percent of all intermediate goods were made in developed countries. At the time, the largest single producer—the United States—accounted for about 20 percentage points of the 70 percent figure. By the 2010s, China's production of intermediate goods surpassed one-quarter of the whole world's production—a figure that is almost twice as large as the next most important supplier (the United States). In 2018, China's manufacturing sector produced a greater value of intermediates than all developed countries combined.

China's rise as a powerhouse of manufactured intermediates production was also rather sudden. At its peak in 2015, China accounted for 42 percent of world manufactured intermediates production, but just ten years earlier, the figure was 14 percent. As shown, the rapid rise has attenuated, and appears to have plateaued, but at a level that implies an astonishing geographic concentration at the world level.

Panel B shows that China's share of global final goods production has been less rapid and less impressive. China's share of world production of final goods and services has also risen compared to 1995 values—seemingly at the expense of developed country production—and is now close to the levels for all other emerging markets. It is, however, still more than 20 percentage points below the collective share of developed nations.

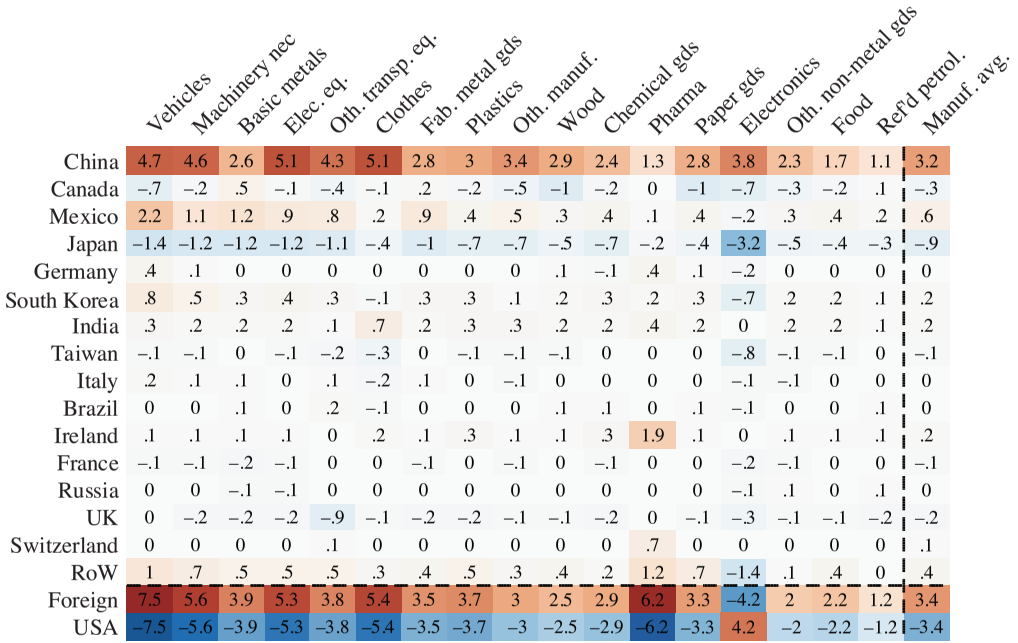
GEOGRAPHIC CONCENTRATION BY SECTOR AND SOURCE NATION China's rise as the premier foreign provider to US supply chains necessarily reduced the relative importance of other suppliers. Further insight into the concentration of US sourcing can be had by looking at the percentage point changes in the shares between 1995 and 2018 by sector and by source nation. Since we are interested in the full impact of the changes, we work with the look-through concept that takes account of all the inputs to the inputs.

Figure 7 displays the numbers, where darker shades of positive numbers indicate higher exposure and darker shades of negative numbers indicate lower exposure in 2018 versus 1995. As in the previous heat maps, it includes the US sourcing from itself. As noted above, the United States, as is true of all mega-economies, supplies most of its own intermediates (as can be seen in the bottom row of figure 4). Figure 7 shows that this self-supplying has diminished. All the entries in the bottom row (the change in the US share of industrial inputs to itself) are negative except for the Electronics sector. The average percentage point (pp) drop across the sectors is 3.4 pp, with the figure ranging from +4.2 pp for the Electronics sector to −7.5 pp in the Vehicles sector. The Pharmaceuticals sector is another standout with a drop of 6.2 pp. The drop in domestic sourcing is matched by an increase in foreign sourcing.

The change in the share provided by all foreign nations is in the next to last row, and these numbers are all positive except in the Electronics column. The most remarkable feature of these numbers is the fact that, apart from Mexico, a large share of the row entries for all the other major suppliers are negative. The simple averages of the changes are only positive for China, Mexico, South Korea, India, Ireland, and Switzerland. China's average change is 3.2 pp, which is far greater than those of the others to which the United States has become more exposed.

It is notable that China's average share rise is only slightly less than the average share drop in US domestic sourcing. In some of the most supply chain-exposed sectors, like Other Transport Equipment and Electrical Equipment, China's percentage point gain is similar to the United States' percentage point drop. The data cannot shed light on how this change occurred, for example, due to offshoring of US intermediate goods

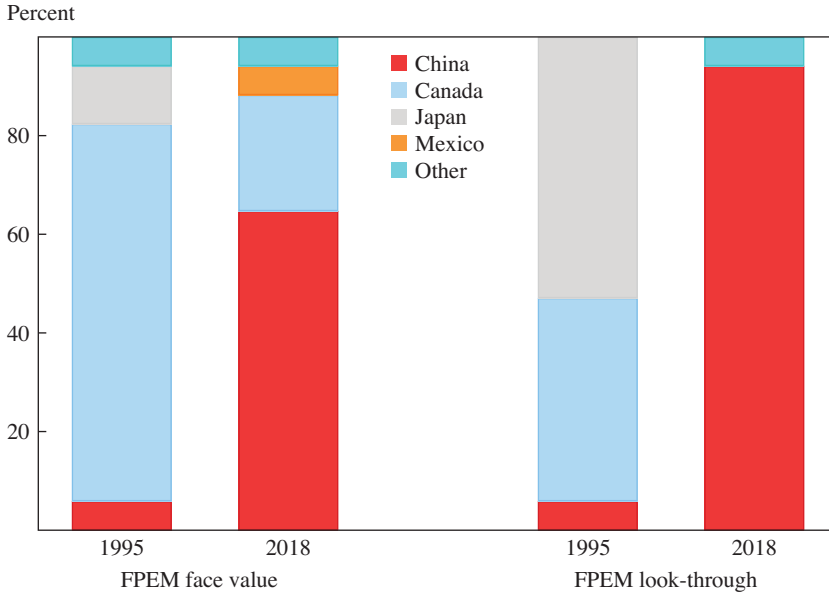
Figure 7. US Look-Through Exposure by Sector (Percentage Point Differences), 1995 versus 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: This figure presents the percentage point difference between look-through exposure in 2018 and look-through exposure in 1995. Look-through exposure is the foreign production exposure: import side (FPEM) indicator as described in Baldwin, Freeman, and Theodorakopoulos (2022) and is computed as the share of manufactured inputs sourced by a given US sector from a given country on a look-through basis in total manufactured intermediates across all sources (foreign and domestic) on a look-through basis. "RoW" stands for the rest of the world. "Foreign" is the sum of all foreign sources.

Figure 8. Top Foreign Supplier of Industrial Inputs to US Manufacturing Sectors, 1995 versus 2018, Face Value versus Look-Through



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: This figure shows the share of US manufacturing sectors for which the top supplier is China, Canada, Japan, Mexico, or other. "FPEM" stands for foreign production exposure: import side (Baldwin, Freeman, and Theodorakopoulos 2022).

production to China, US deindustrialization, or Chinese industrialization. In other sectors, such as Vehicles, the US decline is significantly greater than the Chinese rise since the supply chain also spread to other foreign suppliers. In the Vehicles sector, we see a moderate decline in Canada's and Japan's shares, a big decline in the United States' share, and an important rise in the shares of Mexico, South Korea, and, of course, China.

THE TOP FOREIGN SUPPLIER OF INDUSTRIAL INPUTS OVER TIME The two forms of what we are calling hidden exposure—the look-through versus face value measures on the one hand, and the rapid geographic concentration of sources on the other—can be usefully compared and contrasted by examining the nationality of the top supplier to each of the United States' seventeen manufacturing sectors. Figure 8 shows the share of the seventeen sectors where the top supplier is China, Canada, Mexico, Japan, or some other nation. The chart also shows how this statistic changed from the beginning of our data, 1995, to the end, 2018. The two left-hand columns

use the face value concept to examine the United States' top supplier in 1995 and 2018, while the right-hand columns use the look-through concept in 1995 and 2018.

When it comes to our second form of hidden exposure, the main takeaway from the chart is that China's role as the top supplier spreads rapidly. Turning first to the leftmost pair of stacked columns, we see that in 1995, which was when the new offshoring-oriented globalization was just starting (Baldwin 2006, 2016), China was the top industrial input supplier to about 5 percent of US manufacturing sectors. By 2018, the share was over 60 percent. The change is even starker when using the look-through measure (rightmost pair of stacked columns). China has shifted from being the top supplier in about 5 percent of the sectors to the top supplier in all but one sector (Pharmaceuticals).

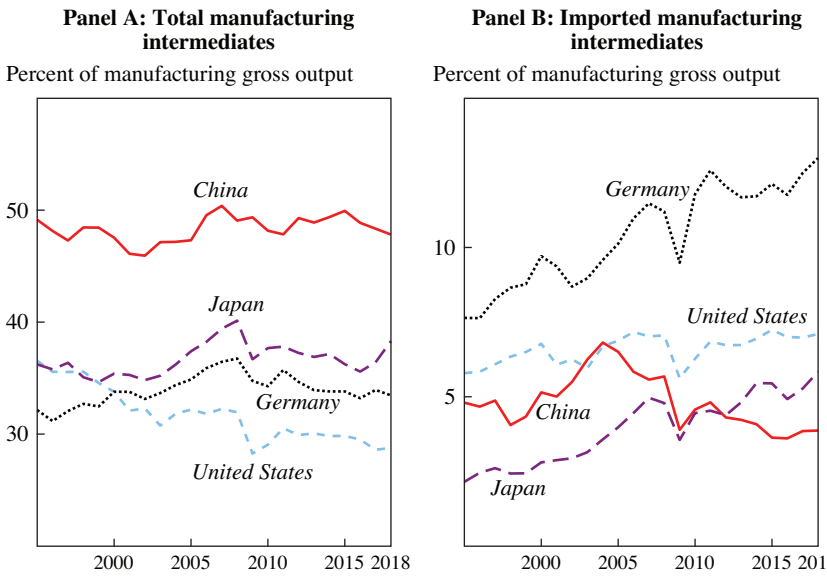
The chart also shows a different take on our first aspect of hidden exposure. Comparing the two stacked columns for 2018 (the second and fourth columns), we see that while China is clearly dominant using the face value concept, it is much more so on a look-through basis.

The chart also illustrates the fact that Japan was, in 1995, playing a similar role to the one that China is playing today. In 1995, the US exposure to foreign industrial inputs was much lower overall since back then the globalization of industrial supply chains was just starting. Most supply chains were domestic. Sticking with the look-through concept to take account of the direct in addition to all indirect sourcing, we see that among the foreign suppliers, Japan had the most top spots. Japan's role, however, looks much less dominant when viewed from the face value perspective. Comparing the first stacked column (1995, face value) to the third stacked column (1995, look-through), we see that the hidden exposure was to Japan back then, not China. This was due to the fact that while the United States was sourcing heavily from Canada, Canada was sourcing heavily from Japan. This was to be expected because Japan was one of the largest producers of intermediate goods outside of the United States.

II.E. Comparison with China

The facts for China could hardly be more different than those for the United States and the two other major manufacturing countries, Germany and Japan. China's industrialization is quite recent compared to that of the United States and other advanced economies, and its development journey was quite different. China started its industrialization with processing trade, which involved limited transformation of imported intermediate goods. From there, China built out its industrial base by producing domestically

Figure 9. Major Manufacturers' Exposure to Supply Chains, 1995–2018



Source: Authors' calculations based on OECD 2021 ICIO tables.

many inputs that had previously been imported. This task was facilitated by its massive and fast-growing internal market and government policy (Cui 2007), foreign investment, and transfers of foreign know-how (Wen 2016). The result is plain to see in figure 9, which also presents the figures for the United States, Japan, and Germany.

Panel A shows the nations' total usage of manufactured intermediates as a share of their manufacturing gross output. We see that Chinese industry is far more exposed to supply chains—taking domestic and international exposure together—than the other three giants. The share of China's manufacturing gross output that is made up of intermediate inputs is about 50 percent, and this figure has been fairly steady since 1995. The corresponding share for the other nations shown is much lower. Panel B, however, shows that Chinese industry is now less exposed to foreign intermediates than the other manufacturing giants. Specifically, China's foreign exposure started in the middle of the pack and rose sharply up to 2005 but has been falling rapidly since. It is now substantially lower in 2018 than that of the others. The United States' exposure to imported manufacturing intermediates is roughly twice, and Germany's is roughly three times that of China.

It is worth noting that all of these "Giant-4" economies are quite self-reliant when it comes to intermediate inputs. The most exposed is Germany, but even there, Germany sources over 85 percent of all its manufacturing intermediates from itself.

Relating this back to figures 4 and 5, we compute that China's average manufacturing look-through exposure to the United States is 0.6 percent (compared to the United States' average manufacturing look-through exposure on China, which is 3.5 percent).[12] And China's average manufacturing hidden exposure to the United States is 0.4, compared to the United States' average manufacturing hidden exposure to China, which is 2.5. Simply put, these counterpart measures underscore that China is much less reliant on US manufacturing than the US manufacturing is on Chinese production.

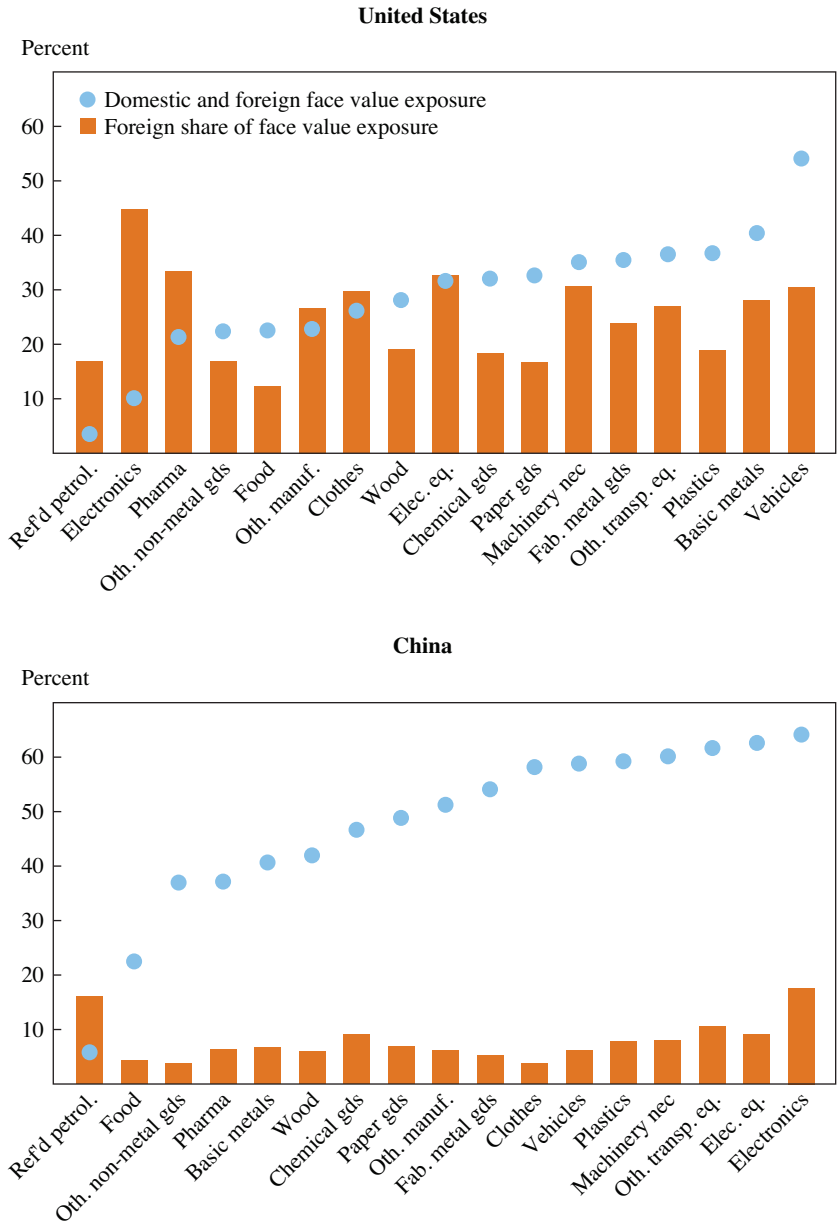
Looking closer, figure 10 shows that China's sectors are generally more exposed overall to supply chains (i.e., combining domestic and foreign sources) but much less exposed to foreign suppliers. For instance, China's foreign exposure is below 20 percent for all sectors, while for the United States, it is much higher, approaching 30 percent to 45 percent in some cases. The opposite holds for the overall (domestic plus foreign) exposure, which is much higher for China than it is for the United States in every single sector.

In terms of geographic concentration, China is also quite different than the United States, as figure 11 shows. This chart, which is comparable to figure 8, shows that China's foreign sourcing is not as concentrated as that of the United States. For instance, the far-right column shows that China's top supplier on a look-through basis is South Korea, but South Korea is the top supplier in only about 60 percent of Chinese manufacturing sectors. Japan, the United States, and other nations play a significant role as top suppliers. On a face value basis (second column from the left), Chinese foreign sourcing is even more diversified.

When it comes to rapid changes in geographic concentration, we do see big upward shifts in South Korea's role from 1995 to 2018, but it is not as stark as the shift that the United States experienced (figure 8). It is also interesting to note that the big hidden exposure for China in 1995 was to Japanese suppliers. On a face value basis (leftmost column), Japan's role was much lower than it was on a look-through basis.

12. Moreover, China's look-through exposure to all nations is comparatively low. Its highest average manufacturing look-through exposure is 1.1 percent with South Korea, which is substantially lower than all other nations' look-through exposure to China.

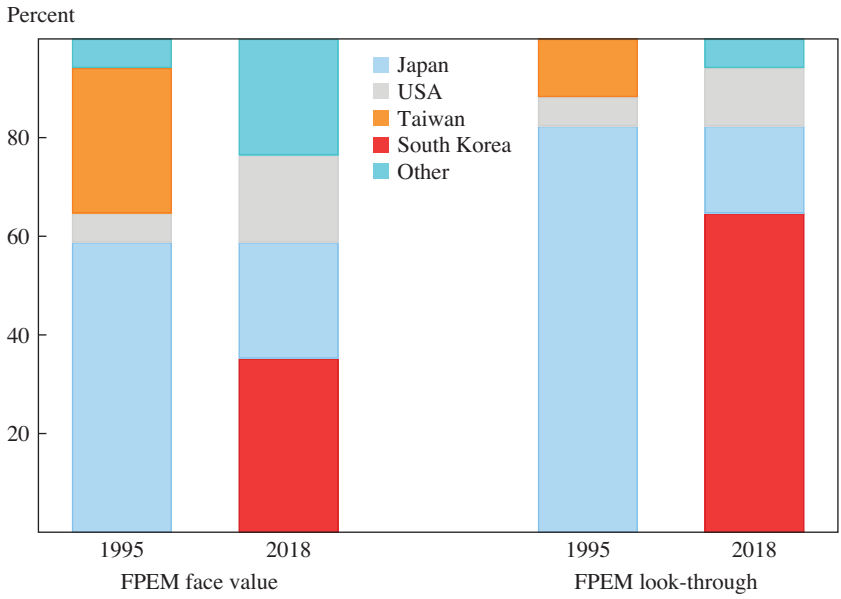
Figure 10. Overall and Foreign Supply Chain Exposure, United States versus China, 2018



Source: Authors' calculations based on OECD 2021 ICIO tables.

Note: This figure shows total (i.e., domestic and foreign) and imported (i.e., foreign) manufacturing intermediate inputs on a face value basis (as a percentage of a sector's gross output). The dots in the United States panel are repeated from figure 2.

Figure 11. Top Foreign Supplier of Industrial Inputs to Chinese Manufacturing Sectors, 1995 versus 2018



Source: Authors' elaboration based on 2021 OECD ICIO tables.

Note: This figure shows the share of Chinese manufacturing sectors for which the top supplier is Japan, South Korea, the United States, Taiwan, or other. "FPEM" stands for foreign production exposure: import side (Baldwin, Freeman, and Theodorakopoulos 2022).

II.F. Measuring Geographic Concentration with Standard Trade Data

The great advantages of IO analysis are the ability to distinguish face value trade from look-through trade and the ability to distinguish between outputs that are used as intermediate goods and those used as final goods. As intermediate goods are what supply chains are set up to acquire, this distinction is critical. The disadvantage that comes with IO analysis is the lack of detail that stems from the very extensive information necessary to estimate the underlying tables, especially at the world level, as opposed to a single-country level.

The sorts of supply chain disruptions that have attracted the attention of heads of state around the world—like those in the semiconductor and medical supply sectors—often involve very specific products. Thus, trade data serve as a valuable complement to the IO analysis since they are available at a much more disaggregated level. The US Census Bureau publishes export and import statistics at the ten-digit level following the

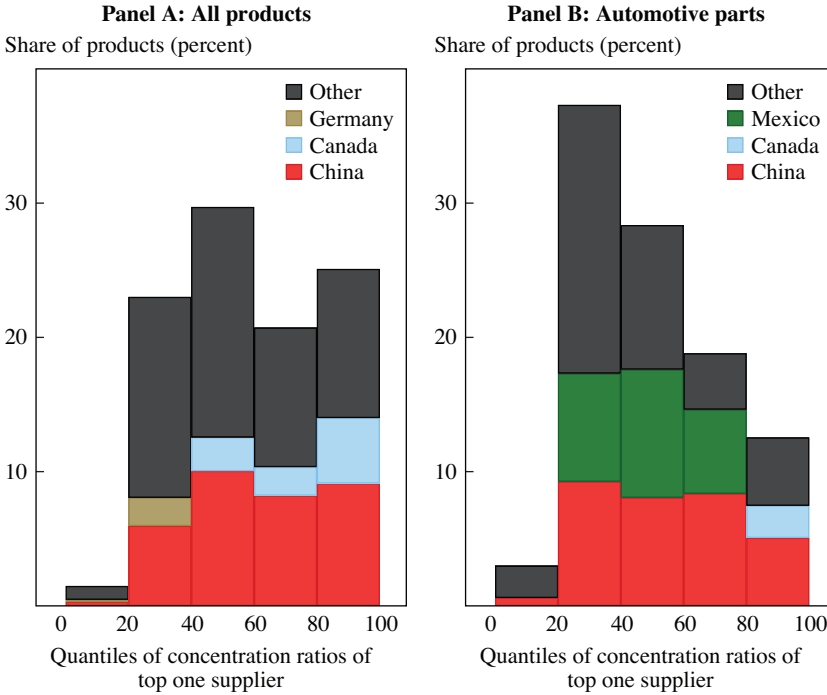
US Harmonized Tariff Schedule (HTS), which distinguishes more than 18,000 different products. To look at the supply chain vulnerability issue from a different perspective, we next turn to the HTS ten-digit data and look for concentration among source nations.

A couple of limitations of the ten-digit data are important to be kept in mind when thinking about the results we will present. The first is that we know neither which sector is importing the goods nor whether they are intermediate or final products. That is, we only know the type of good that is imported into the United States, but we cannot connect the import to a particular purchasing sector. There are some types of imports, like those associated with motor vehicles, where the HTS ten-digit product descriptions allow economists to identify which are intermediate inputs and which are final goods. Moreover, it is reasonable to assume that it is the US auto sector that is purchasing the intermediates. For instance, the product codes 8708305020 for brake drums and 7009100000 for rear-view mirrors are two clear examples. There are other types of imports, such as industrial chemicals, that could be used as inputs in a number of sectors. For these types of imports, we cannot associate geographic concentration with supply chain exposure of a particular sector. As a fallback, we take the exposure as that of the US manufacturing economy as a whole. The second limitation (beyond not always being sure if a product is an intermediate versus final good) is that the trade data only show the face value exposure. For example, if a car part is imported from Canada, we cannot know how much of the good was actually made in Canada and how much was made in another country.

With these caveats in mind, we turn to using the HTS ten-digit trade data to illustrate the geographic concentration of import sourcing. What we look at is the concentration of import sourcing for the 18,043 products that the United States imported in 2018, focusing on imports from a single nation.[13] This is shown in panel A of figure 12; the far-right bar indicates that for about a quarter of all imported products, 80 percent or more of the value came from a single source nation. The bars within the column show the frequency with which the single source supplier is China, Canada, Germany, or some other country. In about a third of the products in this top quintile, the single supplier is China. The other stacked columns in the

13. In line with our concentration analysis with IO data, we focus our attention on a single supplier at the product level. In the absence of firm-level data, we believe that this concentration level reveals particularly high exposure, especially to systemic shocks, which have a broad geographical reach.

Figure 12. Shares of Products Imported by the United States from a Single Source Nation by Quintile of Import Shares, 2018



Source: Authors' elaboration on US Census Bureau trade statistics.

Note: Panel A shows the quintile distribution of all 18,043 products (intermediate and final) imported by the United States in 2018; panel B shows the quintile distribution of all 335 automotive parts (intermediates only) imported by the United States in 2018.

chart are similar, but the bar heights represent goods where the top supplier provides 60–80 percent, 40–60 percent, 20–40 percent, and 0–20 percent of all imports, respectively. Thus, each of the 18,043 products is represented in only one of the five stacked columns.

The first salient fact that emerges from the chart is the remarkable geographic concentration of US imports. The leftmost column indicates that the top supplier was providing less than 20 percent of the total import value for less than 5 percent of the 18,043 products. Considering the two rightmost columns together shows that for almost half the products, more than 60 percent of the import value came from a single supplying nation. In short, the chart indicates a remarkably high level of geographic concentration of import sourcing.

A second noteworthy fact concerns the role of China. In the most concentrated products, for example those underlying the three rightmost columns, China is by far the most important supplier. However, a subtler aspect of this emerges when comparing China's role as a top supplier in figures 8 and 4. We saw in figure 4 that on a look-through basis, China was by far the top supplier in almost every sector. Its dominance is so great that its share of imported inputs was frequently greater than the sum of the next three largest suppliers combined. Yet, panel A in figure 12 would suggest that China is not as dominant a supplier of US imports. For example, for the rightmost column—the one that shows products where at least 80 percent of import value originates from a single nation—China is the top one supplier in only around a third of the cases.

In other words, if one looks at the direct source of imports, China is important but not dominant.[14] However, if one uses IO analysis to determine where the directly imported products were actually made, China's dominant role becomes clear. Of course, the results in figure 4 and figure 12 are not directly comparable, but the contrast is striking. The stark differences are indications of just how much exposure is hidden by failing to look through the veil of inputs into the inputs.[15]

Given the finer level of disaggregation that is possible with trade data, we use the same type of analysis to take a closer look at the United States' imports of automotive parts and components, presumably for the Vehicles sector, where supply chain disruptions are a major issue in the public debate and the distinction between final and intermediate imports is fairly clear. The automotive industry is an interesting case since our IO analysis found it to be one of the most exposed to foreign sourcing, and the nature of automobiles allows us to easily distinguish final from intermediate goods in the HTS ten-digit descriptions. Panel B of figure 12 shows a chart that is similar to the one in panel A, but it focuses solely on the 335 imported products classified as intermediate inputs to the automotive sector by the Office of Transportation and Machinery.[16]

14. Evenett (2020) and Goldberg and Reed (2023) note that face value import dependency from China is small in most product categories.

15. Reconstructing panel A of figure 12 for the top two suppliers (instead of just the top one supplier) reveals that more than half of all the products that the United States imports have over 80 percent of their value coming from just two suppliers.

16. We rely on the US Department of Commerce International Trade Administration classification of automotive parts, as proposed by the Office of Transportation and Machinery, "Harmonized Tariff System Codes, Schedule B Codes, and North American Industry Classification Schedule Codes for Automotive Parts," https://www.trade.gov/automotive-parts-tariff-codes.

A comparison of the two panels of figure 12 suggests that the geographic concentration of supply chain exposure for automotive parts is significantly less marked than it is for the average good (which includes many final goods). The top quintile, for example, covers less than 15 percent of products. This coverage is significantly lower than the 25 percent observed for the entire range of imported goods shown in panel A of figure 9. When considering the top two suppliers, this rises to just over 30 percent. This finding is in line with the findings from figure 5 where we saw that the top six suppliers each provided at least 1 percent of manufactured intermediates to the US Vehicles sector.

III. The Shocks: A Typology of Recent and Likely Future Shocks

To organize thinking and discussions about supply chain shocks, we employ a framework that we proposed in previous work (Baldwin and Freeman 2020b; Baldwin, Freeman, and Theodorakopoulos 2022).[17]

III.A. A Typology of Shocks: Types and Sources

Our typology classifies supply chain shocks into two types—idiosyncratic and systemic—and three sources—supply, demand, and connectivity. The combinations are illustrated with examples in table 1. Supply shocks include classic disruptions such as natural disasters, labor union strikes, the bankruptcy of suppliers, industrial accidents, and the like (Miroudot 2020a). They can also include shocks emanating from broader sources such as trade and industrial policy changes and political instability. Demand shocks can come from many sources. At the firm level, they can be instigated by damage to the reputation of a product or company, customer bankruptcy, entry of new competitors, or policies restricting market access. At the aggregate level, they can be triggered by macroeconomic crises, recessions, or exchange rate changes. Connectivity shocks include, most obviously, transportation of goods, but can also include disruptions of communications and restrictions on travel by key personnel.

The threefold categorization of shock sources is not foolproof. Moreover, one shock can lead to another. The shortage of new US cars and trucks, for example, was a supply shock, but it also created a demand surge that

17. A similar breakdown is also put forth in Goldberg and Reed (2023) in terms of what should be considered when judging responses to economic shocks.

Table 1. Taxonomy of Sources and Nature of Shocks, with Examples

	Supply	*Demand*	*Connectivity*
Idiosyncratic (isolated, simple)	Factory closure, labor strikes, extreme weather, etc.	Single product demand surge, etc.	Single port closure, single firm cyberattack, etc.
Systemic (multi-sector, multi-market, complex interactions)	Pandemics, trade wars, large-scale extreme weather, etc.	Sector-wide preference shifts, multi-product, multi-sector boycotts, embargoes, etc.	Massive hurricanes, military conflicts, large-scale hacking, etc.

Source: Authors' elaboration.

disrupted the used car market (Helper and Soltas 2021). Further, connectivity shocks (such as port congestion and container shortages) can emanate from demand shocks that cause stressed logistics systems or physical disruptions like the Evergreen ship getting stuck in the Suez Canal or reduced traffic in the Panama Canal caused by a severe drought (Doermann 2023). In a similar vein, Guerrieri and others (2022) highlight how COVID-19 started as a supply shock and subsequently led to a demand shock. It is also worth noting that not all shocks fall neatly into the three bins. The destabilizing influences of shifts in trade, taxation, industrial norms, or regulatory guidelines, for example, often defy clear categorization as they can concurrently have an impact on supply, demand, and connectivity.

Importantly, the ability to distinguish among the sources of shocks is crucial, as the appropriate remedies typically depend on identifying the source of the disturbance (Baldwin and Freeman 2022). For example, geodiversifying suppliers will not mitigate unanticipated demand shocks.

SYSTEMIC VERSUS IDIOSYNCRATIC SHOCKS While supply chain disruptions have a long history, we believe that there has been a transformation in the nature of these shocks from mostly idiosyncratic shocks to frequent systemic shocks. And we are not alone. The notion that the nature of shocks shifted is shared by business groups that follow supply chain issues closely (ICC 2023; Hong and Betti 2023).

Leaving aside truly unique events such as the 2008–2009 global financial crisis and the 1970s oil shock, most of the supply chain disruptions before 2016 seemed relatively small, independent, and controllable at the firm level. Notable examples include the floods in Thailand that disrupted auto production, earthquakes in Japan that disrupted the electronics industry, as well as labor strikes. As such, supply chain disruptions seemed to be a topic that could be safely left in the hands of private firms and logistics

companies, supply chain management strategists, and operations research specialists. These shocks were idiosyncratic in nature.

Since 2016, some global supply chains have experienced shocks that affected many sectors and many countries, and some were long-lasting. Notable examples include wars, pandemics, the economic implications of events like Brexit and the US-China geo-economic tensions, and massive cyberattacks like the Colonial Pipeline shutdown discussed below.

This shift in the nature of shocks is a crucial point. Idiosyncratic shocks tend to be controllable at the firm level and thus not an obvious candidate for policy intervention. Systemic shocks, in contrast, are disturbances that resonate across numerous markets, sectors, and products, having a broad geographical and sectoral reach. As such, they are increasingly uncontrollable at the level of individual firms and thus potentially a target for welfare-enhancing policy interventions.

III.B. State of Disruptions

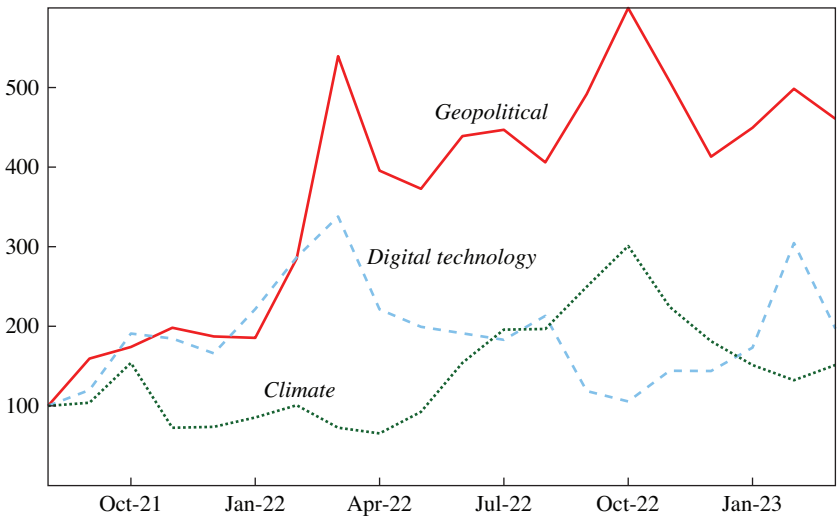
The shocks that emanated from COVID-19 are slowly resolving themselves—at least at the economy-wide level. The Federal Reserve Bank of New York, for example, has developed an index to track the impact of supply chain disruptions (with an eye to their impact on US inflation). This indicator, the Global Supply Chain Pressure Index (GSCPI), spiked at a level that was more than three standard deviations above the historical average in April 2020. The shock faded by October 2020 but then shot up in November 2021 to more than four standard deviations above average. Since then, the GSCPI has fallen.[18] According to the most recent data from July 2023, the GSCPI is nearly a full standard deviation below the index's historical average.[19]

Given that the COVID-19 shocks are fading, it is tempting to think that massive disruptions are a thing of the past and that all the attention being paid to supply chain disruptions by governments and firms is akin to generals preparing for the last war. This is a temptation to resist. While there are no economy-wide data on supply chain disruptions, the COVID-19 shock generated survey-based efforts to gather better and more timely data on shocks. The data, as we shall see, suggest that the supply chain disruption is most definitively not fading, although it is not as intense in 2023 as it was in 2021 and 2022.

18. Federal Reserve Bank of New York, "Global Supply Chain Pressure Index," https://www.newyorkfed.org/research/policy/gscpi.

19. The Bank for International Settlements (BIS), which tracks more classical indicators, comes to roughly the same judgment.

Figure 13. World Economic Forum's Global Value Chain Barometer, 2021–2023



Source: Betti and others (2021), data provided to authors upon request.
Note: Values indexed to 100 in August 2021.

One survey-based gauge, published quarterly by the World Economic Forum (WEF) in partnership with the consulting firm Kearney and utilizing data insights from Everstream Analytics, is the Global Value Chain Barometer (Hong and Betti 2023). In terms of sources of shocks, it focuses on three areas: climate change (especially trade disruptions linked to extreme weather); geo-economic tensions (especially the Russo-Ukrainian War, realignment of emerging-economy coalitions, and trade policy tools that purposefully disrupt trade and investment flows); and digital technologies (especially cybersecurity-related disruption of supply capacities and transportation). After three years of supply chain disruptions driven by the megatrends of climate change, geopolitical tensions, and technological step change, disruption levels seem to have stabilized by 2023:Q1 compared with 2022:Q1, albeit at an elevated level (figure 13). This reflects a combination of a stable trend for new disruptions and firms' improved ability to operate in a more volatile environment. Overall, this suggests that the three big sources of future shocks are not fading in importance.

Another piece of survey evidence regarding current and future supply chain shocks comes from new research by the consulting firm Deloitte, conducted in collaboration with the Federation of German Industries and the

International Service Logistics Association. Their survey, titled "Supply Chain Pulse Check" (Sandau and others 2023), reveals that over half of the supply chain managers from more than 120 German manufacturing enterprises surveyed report a strong to very strong impact on their performance due to supply chain disruptions. A significant majority—60 percent—believe that these disruptions present an even larger problem for the manufacturing sector as a whole. Illustrating the gravity of these supply chain issues—and their potential to worsen—nearly half of the respondents expressed current concerns about a slight to significant increase in the risk of full or partial supply chain failure. These concerned respondents outnumbered those who held the opposite view. Notably, small to medium-sized enterprises indicated a higher level of concern about supply chain disruption and failure compared to large companies.

LIKELY FUTURE SHOCKS Regarding expectations for future shocks, the findings in the Deloitte survey were not optimistic. Almost 60 percent of respondents anticipate no change in the current trend of supply chain disruption, at least in the near-term. Half of them expect a slight improvement in the medium to long term, but over one-fifth foresee the problems becoming slightly or significantly worse in the future.

A similar exercise was undertaken by the Business Continuity Institute (BCI) involving more than 200 supply chain risk management professionals in fifty-eight nations and across seventeen sectors (Elliott, Garcia, and Riglietti 2023). The study found that the reported supply chain disruptions are still more than twice as high as pre-pandemic levels. Almost half of respondents experienced these issues with their closest suppliers at tier 1, while approximately a quarter saw more disruptions with their tier 2 suppliers. Both of these figures exceeded those in the last report (Elliott 2021).[20] Interestingly, the respondents expected cyberattacks and data breaches to be the top threat to supply chains over the coming years.

Looking ahead, the three most cited sources of future systemic shocks are geo-economic tensions, climate change, and digital technology (Hong and Betti 2023; Alicke and others 2022). Geo-economic tensions, for example, have led some actors to use and reshape economic linkages and tools to serve a broader set of strategic goals beyond those that are purely economic, in what some have termed "weaponized interdependence" (Farrell and Newman 2019; Drezner, Farrell, and Newman 2021). For instance, the tariffs implemented by the United States in 2018 were followed by other

20. The 2023 report notes that these high results are partly due to more analysis being undertaken on the performance analytics of supply chains.

countries introducing reciprocal measures to raise trade and investment barriers, often citing geo-economic and national objectives (York 2023; Bown and Kolb 2023). More recently, the Russo-Ukrainian War has not only elevated concerns about supply chains and national security but also triggered a cascade of systemic shocks. These manifest as trade sanctions, boycotts, embargoes, and cross-border restrictions that reverberate through global supply chains, affecting international flows of goods, services, capital, people, and know-how (Goldberg and Reed 2023).

The second source, climate change, is perhaps the ultimate example of radical uncertainty, that is, events whose determinants are insufficiently understood for probabilities to be estimated (King and Kay 2020). Two aspects, however, have clear implications for systemic supply chain disruptions. Extreme weather events have repeatedly knocked production and transportation facilities offline in ways that affect many sectors and many economies (Seneviratne and others 2021). Hurricane Katrina, for example, immediately knocked the Port of New Orleans offline for two weeks and greatly reduced flows through the port for months after. Likewise, heat waves and droughts have forced some electric power plants to reduce output in the United States and France (Barber 2022). On another note, a very different source of shocks concerns future pandemics. Many public health experts expect climate change to induce the migration of species, resulting in novel genetic recombination among animals and thus more zoonotic viruses affecting humans (Randolph and others 2020).

Digital technology is the third source of future systemic shocks. The rapid advance and spread of digital technology in all its manifestations is dialing up the regularity and severity of future shocks in two ways: it is encouraging more activities to shift to the online world where they are vulnerable to accidental and malicious disruptions, and it is boosting the abilities of and incentives for hackers to interrupt normal business activity (Burt 2023). A well-known example is the Colonial Pipeline attack (Easterly and Fanning 2023). In 2021, a criminal hacking group called DarkSide carried out a cyberattack that caused a weeklong disruption in the supply of gasoline to the eastern parts of the United States. The company that owns the pipeline, Colonial Pipeline, had to shut it down to stop the cyber infection and prevent further damage. Since this pipeline was responsible for delivering almost half of the fuel used on the East Coast, the attack led to widespread panic among consumers and a significant increase in fuel prices. Cybersecurity is continuously improving, but so are the skills of criminal and state-sponsored hackers. In this way, digital technology still poses significant risks to supply chain operations around the world.

The last distinction, which is general and applies to all combinations of shocks listed in table 1, is the difference between known unknowns and unknown unknowns. There exists a spectrum of shocks based on our level of awareness and anticipation. At one end are the known unknowns—events or situations we recognize might occur but whose timing and exact form are uncertain. For instance, labor strikes at Paris Charles de Gaulle Airport can be somewhat predicted, given that such events have historical precedents and observable trends. Preparing for these kinds of shocks is relatively straightforward, as we are aware of their potential occurrence. At the opposite end of the spectrum are the unknown unknowns—events without forewarning or precedent and therefore unpredictable in both timing and nature. A fitting example would be the specific characteristics of a future pandemic. While we may anticipate another pandemic based on past occurrences, predicting its exact nature, method of spread, health and economic impacts, and other details is inherently difficult or even impossible.

To provide examples for our policy discussion in the next section, we close this section with a quick recap of recent events before making the case that the nature of supply chain shocks has shifted from idiosyncratic to systemic.

III.C. Brief History of Recent Supply Chain Disruptions

The years 2020–2023 were a wild roller-coaster ride for the world's production networks—a journey into uncharted waters of supply chain bottlenecks, unanticipated dependencies, feedback loops, and formerly hidden interlinkages. But despite the media attention they received, such large-scale supply shocks were not a new thing in 2020. Indeed, who could have imagined, back in early 2019, that the grand challenge to global supply chains would arise from a tiny, malevolent ribbon of RNA?

From 2016, the disruption narrative revolved around geo-economic tensions. These included tariffs imposed by the United States on many of its trade partners and those nations' imposition of retaliatory tariffs (Bown 2017, 2021). The unpredictability of economic policymaking also became a source of disruption. There was also discussion among academics, policymakers, and international organizations about the disruptive possibilities of climate change. These concerns persist today, but their significance was overshadowed by the reach, severity, and lasting impact of the COVID-19 pandemic.

The pandemic took root in late 2019 and surged repeatedly until May 2023 when the World Health Organization officially declared its end as a global health emergency (WHO 2023). A by-product of the disease was

that countries very directly and very expressly disrupted production by imposing stay-at-home measures or reduced-mobility policies that halted factory operations in many sectors worldwide. Other policies also directly disrupted shipping. For example, in an attempt to stall the spread of the virus, many major ports prohibited crew changes without a fourteen-day quarantine, which had a severe impact on transportation and supply chains (Heiland and Ulltveit-Moe 2020; Bai and others 2022).

As nations and businesses were adapting to the virus and related health measures, another source of disruption emerged in 2021. Prevented from spending as much as usual on services like food and entertainment, consumers redirected their expenditures toward physical goods, sparking a resurgence in global demand for manufactured goods. Many such goods were made in Asia or with parts from Asia. This shift in spending patterns intensified disruptions stemming from production and transportation disturbances. The scale and duration of this shift exceeded expectations, and supply struggled to meet surging demand. Critical inputs, such as semiconductors, faced shortages. This had an impact on a range of downstream industries, especially the truck and automobile sectors. The collective effect of these disruptions reveals how fragile and unprepared GSCs were to respond to sudden changes in demand patterns.

An important consequence of this combination of supply and demand shocks was the misplacement of shipping containers due to consumers shifting from in-store to online shopping (Tirschwell 2022). Many of these containers, filled with Asian-manufactured goods, ended up at online fulfillment centers lacking sufficient storage capacity. Furthermore, as the demand surge primarily involved Western demand for goods produced in Asia, trade flows became imbalanced. As containers accumulated in North Atlantic economies, a container shortage emerged in Asia, leading to increased shipping costs and delays. These bottlenecks affected final goods as well as crucial parts and components, ultimately having an impact on manufacturing in the United States and Europe. The pandemic waned and economies reopened in mid-2022, yet global manufacturing remained off-balance. Disruptions persisted due to a near-perfect storm of imbalances. By this, we refer to a convergence of factors—both predictable and unpredictable—that threw supply, demand, and transportation out of equilibrium. The disruptions were so large and so broad that they contributed to an inflationary surge in advanced economies (De Guindos 2023).

The parade of once-in-a-lifetime shocks continued. The Russo-Ukrainian conflict led to sanctions, embargoes, and boycotts, driving commodity and energy prices to soar. This fueled double-digit inflation, which had been

absent for decades, introducing macroeconomic disruptions to production-level shocks. Central banks raised interest rates and global growth slowed. But the surprises did not end there.

A third wave of supply disturbances arose when a new variant of the virus spread to China, triggering severe lockdowns in key centers like Shanghai in the spring of 2022. This hampered shipping and the production of intermediate parts, serving as a less intense but no less significant reminder of the evolving nature of supply chain shocks. Then came China's significant policy reversal—shifting from a stance of zero COVID-19 to almost no policy on COVID-19. After the wave of infections receded, this unleashed pent-up demand from Chinese consumers. China's policy reversal is significant because it not only influences global supply chains but also reveals how quickly governmental policies can change, adding another layer of unpredictability to supply chain planning.

IV. Policy: Robustness and Resiliency

In this section, we explore how the broader, more macroeconomic perspective of the economic approach to supply chains can offer insights that could be valuable in formulating policies to reduce, avoid, or mitigate supply chain disruptions. We start with a critical distinction that is pervasive in the logistics and supply chain management literature (Brandon-Jones and others 2014) but largely absent from the recent economic literature—Miroudot (2020a) is a notable exception—namely, underscoring the difference between robustness and resiliency when it comes to supply chain risk management.

IV.A. Adjusting to Risk: Robustness versus Resilience

Businesses and governments have always been aware of the potential risks of disruption. As the surveys discussed in the previous section showed, firms have put into place adaptive strategies that draw from two vital concepts: robustness and resiliency (Brandon-Jones and others 2014). These words have very similar meanings in English and in fact are sometimes used interchangeably or in tandem in the public discourse surrounding supply chains. To clarify, we start with an example that helps spotlight the differences. The example concerns strategies to address the challenges created by electric power outages.

Most households and businesses understand that the power will occasionally go out and embrace pro-resilience strategies so that they are minimally affected when outages occur. Otherwise stated, they know the shock

will hit and they know operations will be disrupted, but they arrange things to reduce the disruptions and bounce back quickly after the disruption subsides. In contrast, most large hospitals adopt very different strategies, namely, pro-robustness strategies (FEMA 2019). They have multiple alternative electricity sources, including batteries and generators, to ensure that they can continue operating despite the power outage. In a nutshell, the goal of robustness is to have backups that allow the show to go on while the disruption is occurring. The goal of resiliency is to get the show back on the road as soon as possible.

At one level of abstraction, both seek to reduce the duration of production disruptions, but the supply chain risk management literature separates them since the firm-level strategies aimed at robustness are quite different from those aimed at resiliency (Simchi-Levi, Schmidt, and Wei 2014; Simchi-Levi 2015). A supply chain is robust when it continues to operate despite shocks. This is often achieved by engineering supply chains to include fail-safes, redundancies, and geo-diversified supply sources, along with maintaining appropriate inventories of critical inputs. On the sourcing front, robustness signifies cultivating a diversified array of suppliers poised to deliver identical inputs, thereby immunizing the business process against disruptions originating from a single supplier. Within the company's own production sphere, robustness entails maintaining multiple manufacturing sites for in-house inputs and finishing of final goods. In all scenarios, amassing substantial inventory levels and buffer stocks throughout the supply chain, as well as relying on standardized inputs from multiple suppliers, enhances robustness (Sáenz and Revilla 2014).

Resilience relates to the system's capacity for rapid recovery postcrisis, and as such it is a more dynamic concept. The goal is for the supply chain to bounce back from disruptions in a manner that is both efficient and expedient. The essence of resilience lies in flexibility and adaptability, which could take the form of swiftly switching suppliers, adjusting production schedules on the fly, or tweaking products as required (Sá and others 2019; Miroudot 2020b).

Robustness and resilience are not binary options. They are two sides of the same coin in the risk management world. For instance, relying on standardized inputs in a production process (a robustness strategy) could also be a resilience strategy insofar as it would allow flexibility and adaptability in the face of a shock. To summarize, a robust supply chain offers a buffer that can soak up a certain degree of disruption without significant operational impact, buying the system time to respond. In tandem with this,

resilience enables the system to adapt, recover, and thus minimize long-term negative impacts.

TRADE-OFFS IN BUILDING ROBUSTNESS AND RESILIENCE Building robustness and resiliency into supply chains involves distinct sets of strategies. When the shocks come from the supply side, this requires some form of redundancy. This could manifest in a broad and geo-diversified portfolio of suppliers for inputs, multiple production sites, or large inventories. Setting up and maintaining these redundancies necessitates higher immediate operational costs. Indeed, it can be expensive to manage relationships with many suppliers, especially when the input requires extensive checking and certification for quality and fits with the rest of the production process. Further, the spreading out of orders among multiple suppliers may dilute buying power and elevate costs associated with contract supervision and enforcement.

As mentioned, one of the most direct means of establishing robustness is to hold substantial inventories of parts and components, but this can be expensive and even impractical (for example, if warehouse space is not available). One example was the well-anticipated, post-Brexit uncertainty that British carmakers faced when the end of their frictionless trade with the European Union was looming, but they did not really know how well the new system would work. Holding inventory was an obvious idea, but the problem lay in the scale of the challenge. Today's cars are made up of tens of thousands of parts, ranging from nuts and bolts to engines, transmissions, and electronics. Beyond the financial costs of maintaining extensive inventories, the logistical challenge of storing such a wide range of components is formidable.

Moreover, when it comes to highly specialized parts and components, the costs of ensuring that these products meet quality standards and integrate smoothly into the existing production process can make it prohibitively expensive to engage with many suppliers. In such cases, the buyer may have to strive for resiliency rather than robustness. This is why single-sourcing and long-term partnerships often emerge as risk management tactics. While such a strategy might compromise robustness if the supplier encounters risks, the benefits include avoiding the sunk costs of switching suppliers and securing investments from the existing supplier in facilities and practices that can abbreviate disruptions. Even though a serious shock to a single supplier may disrupt overall production, the buyer may choose to put plans in place for quick recovery.

Constructing resilience could involve fostering the ability to adjust production schedules and modify products as required (Miroudot 2020b).

As resilience is likely to involve actions that were not anticipated, off-contract trust among suppliers and buyers (or direct control via ownership) is important in boosting resilience (Sá and others 2019; Dubey and others 2019; Bode and others 2011). In the extreme, resilience may require buyers to functionally control the suppliers or at least maintain long-term relationships that foster sufficient trust. As usual as it is in economics, the choice is not between risk diversification and reliance on lower-cost, higher-quality inputs; it's about finding the right balance. The extra costs today of diversification must be weighed against the expected future benefits of having a supply chain that can carry on in the face of shocks. The possibility that public authorities may have a different evaluation of the trade-off is a key justification for supply chain policy.

IV.B. Do We Need Policy? The Wedge between Private and Public Risk Evaluation

Baldwin and Freeman (2022) introduce an analogy with portfolio theory to discuss the public-private evaluation of supply chain risk. They base this analogy on the standard portfolio model, highlighting the potential existence of a wedge between public and private risk evaluations. While firms are concerned about risks, they also value cost savings. A societal appraisal of this trade-off might prioritize risk reduction more or less highly than the individual firms making the decisions.

EXAMPLES OF PUBLIC-PRIVATE WEDGES IN RISK PERCEPTION What are some examples of these public-private wedges? It is useful to turn to two industries where most governments actively intervene to make the supply chain more resilient: the food sector and the military equipment sector. In the food sector, farmers use various tactics to protect crops from shocks like pests, diseases, and uncertain rainfall. But while the cost to an individual farmer of a bad harvest is limited, a general failure may lead to famine and social upheaval. The wedge here exists because market prices do not fully reflect the social cost of famine or hunger. The classic pro-resiliency government policies in this case are to subsidize production, control prices, and maintain sufficient inventories.

In the realm of military equipment, many governments systematically favor domestic production. While there may be protectionist motives behind such policies, one rationale focuses on the ability to maintain armament production even during wartime. The societal risks associated with a lack of military equipment are even harder to quantify than those in food production. An inability to produce arms and military supplies could lead to loss of territory, loss of life, or loss of sovereignty. In a general way, it is

natural to assume that private firms, which are primarily profit-driven, will underappreciate these social costs of supply disruptions. Protection of basic metals sectors, and steel in particular, is often justified on national security grounds.

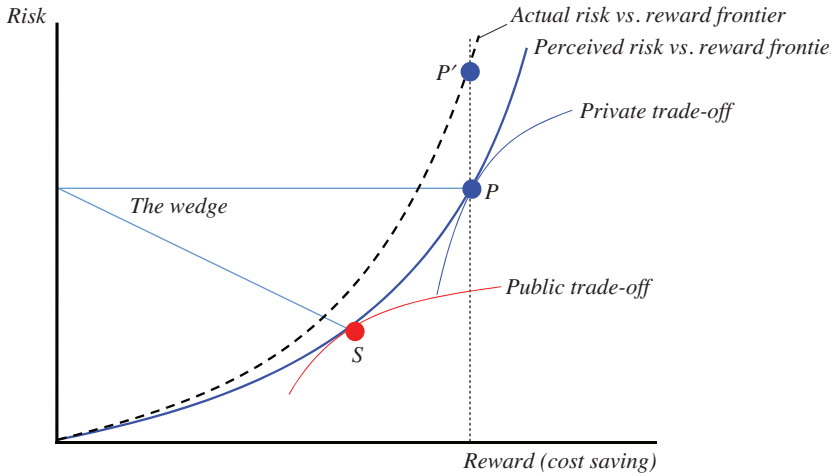
In both the farms and arms cases, we could say that governments knew that the private sector cared about risk, but their caring was mostly limited to their bottom line while the societal cost of major disruptions could be much higher, encompassing factors like social upheaval and loss of life. Another way to rationalize the near-universal intervention of governments in the farms and arms supply chains is the prospect theory of Tversky and Kahneman (1973). This theory explains how humans tend to act in seemingly irrational ways in the face of uncertainty. It stresses the role of present-biased reference points, pervasive loss aversion, and the importance of framing effects.

In the financial sector as well, governments seldom entrust risk management entirely to private entities. The justifications for the interventions are wide-ranging, but many are rooted in information asymmetry, inadequate information, or some agents' inability to process information correctly. These range from investor protection and transparency rules to market stability policies.

Elements of the justifications from these three examples are clear in the recent spate of risk management policies put forth by the Biden administration (White House 2021). The executive order asserts that structural weaknesses in United States supply chains have long existed, but it took the COVID-19 pandemic to bring them into the mainstream. The document notes the need to "strengthen critical supply chains and rebuild [the US] industrial base" (White House 2021, 12). The Biden administration's policy has focused on four sectors that share some of the characteristics of the food and military supply sectors on the one hand, and the financial sector on the other. These are: semiconductors and advanced packaging; large-capacity batteries; critical minerals and materials; and pharmaceuticals and related active pharmaceutical ingredients.

Semiconductors and batteries have become critical to the production of many manufactured goods, including a wide range of armaments. The justification for public policies may thus be linked to those that apply to the arms industry. The advanced packaging concern came to light when the US rollout of COVID-19 vaccines was delayed by a lack of glass vials with the necessary quality. The inclusion of pharmaceuticals can be thought of as akin to the justifications for intervention in the food sector. While individual producers are aware of risks and take active measures to reduce

Figure 14. The Risk-Reward Wedge and Public Policy



Source: Baldwin and Freeman (2022), adapted with permission from the *Annual Review of Economics* 14, copyright 2022 by Annual Reviews.

them, they do not fully incorporate the social costs of severe supply shortages into their business models.

THE WEDGE DIAGRAM Every economics student learns that policy interventions can potentially rectify market outcomes when there is a wedge between the private and public evaluation of the consequences. This happens when there is a gap between private and societal risk assessments, or when collective action challenges cause information gaps, leading firms to operate without full information. Figure 14, presented in Baldwin and Freeman (2022), illustrates these points.

The central idea that the diagram illustrates concerns a trade-off between cost savings and risk. That is, firms can lower costs by centralizing production in cost-efficient areas. However, this cost-saving approach increases the risk associated with centralizing all production. The diagram illustrates this trade-off with the upward-sloped risk-reward curve that is bowed outward. This curve simply asserts that additional cost savings come with heightened risks. The risk-reward frontier curves upward, indicating that the risks-versus-cost-savings trade-off steepens as costs fall.

The downward-curving private-evaluation curve is an indifference curve. It reflects the trade-off firms face on the economic side. That is to say, while firms dislike risk, they like cost savings. The "Private trade-off" curve depicts this relative evaluation. This indifference curve is bowed

downward since we assume that firms worry more about risk as the risk level rises. In other words, firms need ever greater increments in cost savings to justify ever higher risk.

The diagram also plots the public evaluation of the risk-reward tradeoff, which is drawn assuming that the government is more risk-averse than private firms. Various reasons—such as those discussed above in the farms and arms sectors—can justify this discrepancy. For instance, companies might overlook the broader macroeconomic ramifications of supply disruptions, focusing solely on their own performance. Disruptions at one supply chain point could result in losses downstream, but upstream entities might not factor in these potential losses.

As mentioned, such a gap between public and private risk perceptions is easy to envision in critical sectors like medical supplies or food production or for other strategic inputs like semiconductors. As illustrated in figure 14, private entities, in the pursuit of their private goals, might be willing to embrace more risk (as shown by point *P*) than would be socially optimal (point *S*). This difference between societal and private preferences creates a discernible gap and hence a market inefficiency. This inefficiency, in turn, suggests a rationale for policy interventions that reduce supply chain risk.

Importantly, this wedge is not a classic Pigouvian wedge that arises from divergences between public and private evaluations of the marginal benefits or marginal costs of an activity. Our wedge is based on risk perceptions. As drawn, the government is more worried about risk than the private sector, but it could plausibly go the other way. For example, the government would want the firm to take more risk with its supply chain in order to accelerate the delivery of, say, vaccines to the market.

The diagram also sheds light on another possible reason for policy action: information problems. As discussed in section II, firms often have incomplete information about their supply chains due to their sheer complexity. The McKinsey Global Institute's estimate that General Motors had over 18,000 suppliers serves as a telling example; monitoring all these suppliers would be nearly impossible (Lund and others 2020). Moreover, the same study found that nearly half of the companies that were assessed either had no detailed information on their supply chains or had information only on their immediate, tier 1 suppliers. With such a complex web of suppliers, it's hardly surprising that firms may inadvertently expose themselves to more risks than they assume. In other words, the actual risk landscape might be far more perilous than perceived, leading firms to make choices that unknowingly expose them to undue risks.

V. Concluding Remarks

Our paper looks at the three fundamental elements of supply chain disruptions: the links that create the possibility of disruption; the shocks that create the disruptions; and measures aimed at taming or avoiding the disruptions. Here in the concluding remarks, we put forward some conjectures on the implications of our discussion of the three elements.

Starting with the links element, a core message of our paper is that the United States' exposure to foreign supply chains is much bigger than it appears at face value, but it is not that big at the macro level. There are two distinct points in this bigger-but-not-big finding.

First, by any measure, the United States buys at least 80 percent of all industrial inputs from domestic sources. Thus, at an aggregate level, its foreign exposure is hardly alarming. However, while this may be reassuring, it is important to note that supply chain disruptions rarely occur at the macro level. The 80 percent figure was not relevant when the US auto sector shuttered factories due to a lack of semiconductors, or when buying home office electronics became problematic due to a demand surge and logistic snarls. This observation serves to provide some perspective on the recent public debate on foreign supply chains. Concerns about foreign exposure should be directed to particular products, not US manufacturing as a whole (more on this below). This is our conjecture as to what the "not big" part of our results means. The bigger part of bigger-but-not-big suggests a very different conjecture.

US supply chain exposure to some foreign suppliers is much higher than it appears to be using standard trade statistics. We calculate that this is especially true for China. By any measure, China is the United States' largest supplier of industrial inputs. But taking account of the Chinese inputs into all the inputs that American manufacturers buy from other foreign suppliers—what we call look-through exposure—we see that the US exposure to China is almost four times larger than it appears to be at face value.[21] A second aspect of hidden exposure arises from the fact that China's dominance of the United States' imports of industrial inputs came rather suddenly. This might help explain why the basic point was not brought to the fore until recently.

21. The same hidden exposure point holds for Taiwan and South Korea. Their look-through exposure is 3.5 times larger. For Japan the figure is 3.1. Nonetheless, these countries have a much smaller absolute face value and look-through exposure overall.

Combining the two points from our links results, in conjunction with the fact that all major economies are also highly reliant on Chinese inputs to their inputs suggests that an across-the-board decoupling of the US and Chinese manufacturing sectors is unlikely to be cheap, quick, or even feasible.[22] More research is needed to quantify this point, but recent studies all point to the fact that a US-China decoupling is likely to be very damaging economically to the United States and the world as a whole (Góes and Bekkers 2021; Freund and others 2023; Métivier and others 2023; Aiyar and others 2023).

Moreover, taking the face value versus look-through distinction to heart suggests that the latter measure is more relevant in assessing whether policies aimed at reducing US exposure to Chinese manufacturing will have their desired effect. For instance, simply substituting away from imports from China to, say, Vietnam may do little to reduce the look-through dependence on Chinese production if the Vietnamese exports to the United States depend on Chinese inputs. This important point is made empirically by Alfaro and Chor (2023).

Turning to the second element of supply chain disruptions, the shocks, our discussion suggests that the United States is facing a new reality when it comes to supply chain shocks. We argue that the nature of shocks has shifted. While idiosyncratic shocks continue to produce challenges for manufacturers around the world, many of the recent and likely future shocks will be systemic. Here idiosyncratic shocks are those that are isolated and limited in scope, while systemic shocks have impacts that affect multiple sectors and regions and may be long-lasting. In addition to these two types of shocks, we underscore that the source of supply chain shocks can be either demand-driven, supply-driven, or affect connectivity—and that these three categories are often interconnected.

While there is no way to predict future shocks—and in particular those that are systemic in nature—evidence gathered from surveys of supply chain risk managers coupled with the costly, long-lasting adjustments that firms are making to their supply chain organization, provides evidence that the nature of shocks has shifted. These surveys highlighted three central sources of future shocks: climate change, geo-economic tensions, and accidental and malicious digital disruptions.

22. As mentioned in section II.C., our look-through measure also tells us that it is not just the United States that is heavily dependent on China for industrial supplies; every major manufacturing nation in the world sources at least 2 percent of its industrial intermediates from China (Baldwin, Freeman, and Theodorakopoulos 2022).

Laying our findings on shocks end to end with our findings on links leads to a very clear policy message. Concerns about supply chain disruptions should not be overblown, but they should be taken seriously since they are likely to be with us for many years to come.

The final element of our paper concerns policies that are aimed at reducing the impact of supply chain disruptions. As an essential background to policy considerations, we highlighted here the need to think hard about rationales for public policy interventions. A second bit of essential background that we touched upon is the nontrivial distinction between robustness and resiliency in supply chains, which is taken as critical in supply chain risk management research. The need for a policy intervention rationale is clear, but we focus on divergence in the evaluation of risk by the government and private sector, not the traditional situation that focuses on market inefficiencies.

Because firms actively choose the risk level of their supply chains (to the extent that they have visibility of their suppliers and suppliers' suppliers), any public policy intervention should be based on the presence of a public-private wedge in the trade-off between cost savings and disruption risk. Given the vast diversity in supply chains, we argued this point by analogy, drawing attention to sectors where most nations have chosen to interfere with the private sector's optimal combination of low-cost sourcing and concentration of supply chain risk. In the farms and arms sectors, for example, governments have long implemented expensive policy interventions to encourage domestic production and diversified sources. In these sectors, the public-private wedge arises from many underlying factors, but often they involve the fact that serious disruptions can create large-scale societal problems. As the private sector has little incentive to fully internalize such problems, it is easy to imagine that the wedge is large in these sectors.

Do the sectors that have recently been the focus of government supply chain policy fit this bill? In the United States, Europe, and Asia, semiconductors seem to have slipped into the same category as farms and arms in the sense that governments around the world have decided that they cannot rely solely on the private sector to control supply chain risks. In the United States, the Biden administration has also put some pharmaceutical products as well as large-capacity batteries into the farms and arms category. Without detailed simulations of the economic and social costs of disruptions in these products, it is impossible to comment precisely on the merit of these governmental choices. But, given the lack of incentives for

firms to consider the broader societal costs of extreme events, it is easy to think that there are wedges that would justify intervention in these sectors.

V.A. Directions for Future Research

It is plain that there is much, much more that could be done to shed light on the exposure of US supply chains to future shocks. One direction would be to explore the use of granular data, such as firm-specific, transaction-level data or fine-grained geographic data.[23] In particular, it would be very helpful to have more disaggregated ICIO tables at the country and industry dimensions to gain a deeper understanding of supply chain vulnerabilities and the propagation of disruptions in further detail. It would also be useful to more fully document how supply chain exposure became so concentrated geographically. Adding econometric investigation would also be an important contribution. The OECD, for example, has used some of the look-through measures we developed in our earlier work to demonstrate that they provide a more robust empirical accounting for the transmission of shocks than do face value measures (Schwellnus, Haramboure, and Samek 2023a; Schwellnus and others 2023b). The last point we mention is the extension of the entire face value versus look-through distinction to an evaluation of the exposure of US manufacturing sectors on the sales side, that is to say, the exporting side.

ACKNOWLEDGMENTS We would like to thank Janice Eberly, Jón Steinsson, Pinelopi Goldberg, and Benjamin Golub for helpful comments on an earlier draft of this paper that resulted in many improvements. We would also like to thank Thomas Prayer for assistance downloading US trade data from the US Census Bureau.

23. To be sure, additional measures of supply chain exposure are being developed, in particular using product-level data. For instance, concerned primarily with the possibility of supply disruptions, the European Commission (2021) and Arjona, García, and Herghelegiu (2023) recently proposed a methodology for measuring the European Union's strategic dependencies and vulnerabilities at the detailed product level, which relies on the computation and use of three indicators relating to the concentration of EU imports from non-EU sources, the importance of non-EU imports in total demand, and the substitutability of non-EU imports with EU production.

References

Aiyar, Shekhar, Jiaqian Chen, Christian H. Ebeke, Roberto Garcia-Saltos, Tryggvi Gudmundsson, Anna Ilyina, and others. 2023. "Geoeconomic Fragmentation and the Future of Multilateralism." Staff Discussion Notes 2023/001. Washington: International Monetary Fund.

Alfaro, Laura, and Davin Chor. 2023. "Global Supply Chains: The Looming 'Great Reallocation.'" In *Economic Policy Symposium Proceedings: Structural Shifts in the Global Economy*. Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Alicke, Knut, Edward Barriball, Tacy Foster, Julien Mauhourat, and Vera Trautwein. 2022. "Taking the Pulse of Shifting Supply Chains." New York: McKinsey Global Institute. https://www.mckinsey.com/capabilities/operations/our-insights/taking-the-pulse-of-shifting-supply-chains.

Antràs, Pol. 2020. "De-Globalisation? Global Value Chains in the Post-COVID-19 Age." Working Paper 28115. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w28115.

Antràs, Pol, and Davin Chor. 2022. "Global Value Chains." *Handbook of International Economics, Volume 5*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Amsterdam: North-Holland.

Arjona, Román, Wiliam Connell García, and Cristina Herghelegiu. 2023. "An Enhanced Methodology to Monitor the EU's Strategic Dependencies and Vulnerabilities." Working Paper 14. Brussels: European Commission.

Bai, Xiwen, Ming Xu, Tingting Han, and Dong Yang. 2022. "Quantifying the Impact of Pandemic Lockdown Policies on Global Port Calls." *Transportation Research Part A: Policy and Practice* 164: 224–41.

Baldwin, Richard. 2006. *Globalisation: The Great Unbundling(s)*. Helsinki: Economic Council of Finland.

Baldwin, Richard. 2016. *The Great Convergence: Information Technology and the New Globalization.* Cambridge, Mass.: Harvard University Press.

Baldwin, Richard. 2022. "The Peak Globalisation Myth: Part 1." VoxEU, August 31. https://cepr.org/voxeu/columns/peak-globalisation-myth-part-1.

Baldwin, Richard, and Rebecca Freeman. 2020a. "Supply Chain Contagion Waves: Thinking Ahead on Manufacturing 'Contagion and Reinfection' from the COVID Concussion." VoxEU, April 1. https://cepr.org/voxeu/columns/supply-chain-contagion-waves-thinking-ahead-manufacturing-contagion-and-reinfection.

Baldwin, Richard, and Rebecca Freeman. 2020b. "Trade Conflict in the Age of Covid-19." VoxEU, May 22. https://cepr.org/voxeu/columns/trade-conflict-age-covid-19.

Baldwin, Richard, and Rebecca Freeman. 2022. "Risks and Global Supply Chains: What We Know and What We Need to Know." *Annual Review of Economics* 14:153–80.

Baldwin, Richard, Rebecca Freeman, and Angelos Theodorakopoulos. 2022. "Horses for Courses: Measuring Foreign Supply Chain Exposure." Working

Paper 30525. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30525.

Baldwin, Richard, Rebecca Freeman, and Angelos Theodorakopoulos. 2023. "Deconstructing Deglobalization: The Future of Trade Is in Intermediate Services." *Asian Economic Policy Review* 19, no. 1: 18–37.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87, no. 4: 1155–203.

Baqaee, David Rezza, and Elisa Rubbo. 2023. "Micro Propagation and Macro Aggregation." *Annual Review of Economics* 15: 91–123.

Barber, Gregory. 2022. "Nuclear Power Plants Are Struggling to Stay Cool." *Wired*, July 21. https://www.wired.com/story/nuclear-power-plants-struggling-to-stay-cool/.

Betti, Francisco, Felipe Bezarnat, Mernia Fendri, Benjamin Henkes, Per Kristian Hong, and Xavier Mesnard. 2021. "The Resiliency Compass: Navigating Global Value Chain Disruption in an Age of Uncertainty." Geneva: World Economic Forum.

Bode, Christoph, Stephan M. Wagner, Kenneth J. Petersen, and Lisa M. Ellram. 2011. "Understanding Responses to Supply Chain Disruptions: Insights from Information Processing and Resource Dependence Perspectives." *Academy of Management Journal* 54, no. 4: 833–56.

Boehm, Christoph E., Aaron Flaaen, and Nitya Pandalai-Nayar. 2019. "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake." *Review of Economics and Statistics* 101, no. 1: 60–75.

Bonadio, Barthélémy, Zhen Huo, Andrei A. Levchenko, and Nitya Pandalai-Nayar. 2021. "Global Supply Chains in the Pandemic." *Journal of International Economics* 133:103534.

Bown, Chad P. 2017. "Trump's Attempted Takedown of the Global Trade Regime?" In *The Future of the Global Order Colloquium: Fall 2017*. Philadelphia: Perry World House, University of Pennsylvania.

Bown, Chad P. 2021. "The US-China Trade War and Phase One Agreement." *Journal of Policy Modeling* 43, no. 4: 805–43.

Bown, Chad P., and Melina Kolb. 2023. "Trump's Trade War Timeline: An Up-to-Date Guide." Blog post, April 19, 2018, Peterson Institute for International Economics. https://www.piie.com/blogs/trade-and-investment-policy-watch/2018/trumps-trade-war-timeline-date-guide.

Brandon-Jones, Emma, Brian Squire, Chad W. Autry, and Kenneth J. Petersen. 2014. "A Contingent Resource-Based Perspective of Supply Chain Resilience and Robustness." *Journal of Supply Chain Management* 50, no. 3: 55–73.

Burt, Andrew. 2023. "The Digital World Is Changing Rapidly. Your Cybersecurity Needs to Keep Up." *Harvard Business Review*, May 16. https://hbr.org/2023/05/the-digital-world-is-changing-rapidly-your-cybersecurity-needs-to-keep-up.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi. 2021. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." *Quarterly Journal of Economics* 136, no. 2: 1255–321.

Carvalho, Vasco M., and Alireza Tahbaz-Salehi. 2019. "Production Networks: A Primer." *Annual Review of Economics* 11: 635–63.

Conconi, Paola, Glenn Magerman, and Afrola Plaku. 2020. "The Gravity of Intermediate Goods." *Review of Industrial Organization* 57: 223–43.

Council of Economic Advisers (CEA). 2016. *Economic Report of the President*. Washington: US Government Publishing Office. https://www.govinfo.gov/content/pkg/ERP-2016/pdf/ERP-2016-frontmatter.pdf.

Cui, Li. 2007. "China's Growing External Dependence." *Finance and Development* 44, no. 3: 42–45.

De Guindos, Luis. 2023. "The Inflation Outlook and Monetary Policy in the Euro Area." Keynote speech at King's College London, July 7. https://www.ecb.europa.eu/press/key/date/2023/html/ecb.sp230707~8f8f9debc6.en.html.

Doermann, Lindsey. 2023. "Panama Canal Traffic Backup." NASA Earth Observatory, August 18. https://earthobservatory.nasa.gov/images/151778/panama-canal-traffic-backup.

Drezner, Daniel W., Henry Farrell, and Abraham L. Newman, eds. 2021. *The Uses and Abuses of Weaponized Interdependence*. Washington: Brookings Institution Press.

Dubey, Rameshwar, Angappa Gunasekaran, Stephen J. Childe, Thanos Papadopoulos, Constantin Blome, and Zongwei Luo. 2019. "Antecedents of Resilient Supply Chains: An Empirical Study." *IEEE Transactions on Engineering Management* 66, no. 1: 8–19.

Easterly, Jen, and Tom Fanning. 2023. "The Attack on Colonial Pipeline: What We've Learned and What We've Done over the Past Two Years." Blog post, May 7, US Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/news-events/news/attack-colonial-pipeline-what-weve-learned-what-weve-done-over-past-two-years.

Elliott, Matthew, and Benjamin Golub. 2022. "Networks and Economic Fragility." *Annual Review of Economics* 14: 665–96.

Elliott, Matthew, Benjamin Golub, and Matthew V. Leduc. 2022. "Supply Network Formation and Fragility." *American Economic Review* 112, no. 8: 2701–47.

Elliott, Rachael. 2021. *Supply Chain Resilience Report 2021*. Caversham: Business Continuity Institute.

Elliott, Rachael, Maria Florencia Lombardero Garcia, and Gianluca Riglietti. 2023. *BCI Supply Chain Resilience Report 2023*. Caversham: Business Continuity Institute.

European Commission. 2021. "Strategic Dependencies and Capacities." Commission Staff Working Document. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021SC0352.

European Council. 2023. "G7 Leaders' Statement on Economic Resilience and Economic Security." Press release, May 20. https://www.consilium.europa.eu/en/press/press-releases/2023/05/20/g7-leaders-statement-on-economic-resilience-and-economic-security/.

Evenett, Simon J. 2020. "Chinese Whispers: COVID-19, Global Supply Chains in Essential Goods, and Public Policy." *Journal of International Business Policy* 3: 408–29.

Evenett, Simon J. 2021. *Trade Policy and Medical Supplies during COVID-19: Ideas for Avoiding Shortages and Ensuring Continuity of Trade*. London: Chatham House.

Farrell, Henry, and Abraham L. Newman. 2019. "Weaponized Interdependence: How Global Economic Networks Shape State Coercion." *International Security* 44, no. 1: 42–79.

Federal Emergency Management Agency (FEMA). 2019. *Healthcare Facilities and Power Outages: Guidance for State, Local, Tribal, Territorial, and Private Sector Partners*. Washington: Author.

Feenstra, Robert C. 2009. *Offshoring in the Global Economy: Microeconomic Structure and Macroeconomic Implications*. Cambridge, Mass.: MIT Press.

Fort, Teresa C. 2023. "The Changing Firm and Country Boundaries of US Manufacturers in Global Value Chains." *Journal of Economic Perspectives* 37, no. 3: 31–58.

Freund, Caroline, Aaditya Mattoo, Alen Mulabdic, and Michele Ruta. 2023. "Is U.S. Trade Policy Reshaping Global Supply Chains?" Policy Research Working Paper 10593. Washington: World Bank. https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-10593.

Góes, Carlos, and Eddy Bekkers. 2021. "The Impact of Geopolitical Conflicts on Trade, Growth, and Innovation." Staff Working Paper ERSD-2022-09. Geneva: World Trade Organization.https://www.wto.org/english/res_e/reser_e/ersd202209_e.htm.

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

Grossman, Gene M., Elhanan Helpman, and Hugo Lhuillier. 2021. "Supply Chain Resilience: Should Policy Promote Diversification or Reshoring?" Working Paper 29330. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w29330.

Grossman, Gene M., and Esteban Rossi-Hansberg. 2008. "Trading Tasks: A Simple Theory of Offshoring." *American Economic Review* 98, no. 5: 1978–97.

Guerrieri, Veronica, Guido Lorenzoni, Ludwig Straub, and Iván Werning. 2022. "Macroeconomic Implications of COVID-19: Can Negative Supply Shocks Cause Demand Shortages?" *American Economic Review* 112, no. 5: 1437–74.

Gurtu, Amulya, and Jestin Johny. 2021. "Supply Chain Risk Management: Literature Review." *Risks* 9, no. 1: 1–16.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse, Toolkit, and Cookbook." In *Handbook of International Economics, Volume 4*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Amsterdam: North-Holland.

Heiland, Inga, and Karen Helene Ulltveit-Moe. 2020. "An Unintended Crisis in Sea Transportation Due to COVID-19 Restrictions." In *COVID-19 and Trade Policy: Why Turning Inward Won't Work*, edited by Richard E. Baldwin and Simon J. Evenett. London: Centre for Economic Policy Research.

Helper, Susan, and Evan Soltas. 2021. "Why the Pandemic Has Disrupted Supply Chains." Blog post, June 17, The White House Council of Economic Advisers. https://www.whitehouse.gov/cea/written-materials/2021/06/17/why-the-pandemic-has-disrupted-supply-chains/.

Hong, Per Kristian, and Francisco Betti. 2023. "Navigating Global Disruption: Introducing the Global Value Chain Barometer." World Economic Forum, January 13. https://www.weforum.org/agenda/2023/01/davos23-global-value-chain-barometer/.

Houthakker, Hendrik S. 1955. "The Pareto Distribution and the Cobb-Douglas Production Function in Activity Analysis." *Review of Economic Studies* 23, no. 1: 27–31.

Imbs, Jean, and Laurent Pauwels. 2022. "Measuring Openness." Discussion Paper 17230. London: Centre for Economic Policy Research.

International Chamber of Commerce (ICC). 2023. *ICC 2023 Trade Report: A Fragmenting World*. Paris: Author.

Jones, Charles I. 2005. "The Shape of Production Functions and the Direction of Technical Change." *Quarterly Journal of Economics* 120, no. 2: 517–49.

King, Mervyn, and John Kay. 2020. *Radical Uncertainty: Decision-Making for an Unknowable Future.* New York: W. W. Norton & Company.

Lund, Susan, James Manyika, Jonathan Woetzel, Ed Barriball, Mekala Krishnan, Knut Alicke, and others. 2020. *Risk, Resilience, and Rebalancing in Global Value Chains*. New York: McKinsey Global Institute.

Métivier, Jeanne, Marc Bacchetta, Eddy Bekkers, and Robert Koopman. 2023. "International Trade Cooperation's Impact on the World Economy." Staff Working Paper ERSD-2023-02. Geneva: World Trade Organization. https://www.wto.org/english/res_e/reser_e/ersd202302_e.pdf.

Miroudot, Sébastien. 2020a. "Reshaping the Policy Debate on the Implications of COVID-19 for Global Supply Chains." *Journal of International Business Policy* 3: 430–42.

Miroudot, Sébastien. 2020b. "Resilience versus Robustness in Global Value Chains: Some Policy Implications." In *COVID-19 and Trade Policy: Why Turning Inward Won't Work*, edited by Richard E. Baldwin and Simon J. Evenett. London: Centre for Economic Policy Research.

Miroudot, Sébastien., Rainer Lanz, and Alexandros Ragoussis. 2009. "Trade in Intermediate Goods and Services." OECD Trade Policy Paper 93. Paris: Organisation for Economic Co-operation and Development.

Moll, Benjamin, Moritz Schularick, and Georg Zachmann. 2023. "The Power of Substitution: The Great German Gas Debate in Retrospect." In the present volume of *Brookings Papers on Economic Activity*.

Mourougane, Annabelle, Polina Knutsson, Rodrigo Pazos, Julia Schmidt, and Francesco Palermo. 2023. "Nowcasting Trade in Value Added Indicators."

OECD Statistics Working Paper 2023/03. Paris: Organisation for Economic Co-operation and Development. https://www.oecd.org/sdd/nowcasting-trade-in-value-added-indicators-00f8aff7-en.htm.

Oberfield, Ezra, and Devesh Raval. 2021. "Micro Data and Macro Technology." *Econometrica* 89, no. 2D: 703–32.

Porter, Michael E. 1985. *Competitive Advantage: Creating and Sustaining Superior Performance*. New York: Free Press.

Randolph, Della Grace, Johannes Refisch, Susan MacMillan, Caradee Yael Wright, Bernard Bett, Doreen Robinson, and others. 2020. *Preventing the Next Pandemic: Zoonotic Diseases and How to Break the Chain of Transmission*. Nairobi: United Nations Environment Programme.

Sá, Marcelo Martins de, Priscila Laczynski de Souza Miguel, Renata Peregrino de Brito, and Susana Carla Farias Pereira. 2019. "Supply Chain Resilience: The Whole Is Not the Sum of the Parts." *International Journal of Operations and Production Management* 40, no. 1: 92–115.

Sáenz, María Jesús, and Elena Revilla. 2014. "Creating More Resilient Supply Chains." *MIT Sloan Management Review*, June 17. https://sloanreview.mit.edu/article/creating-more-resilient-supply-chains/.

Sandau, Jürgen, Yasmin Wächter, Alexander Börsch, and Florian Ploner. 2023. *Supply Chain Pulse Check: New Risks for the Supply Chain and the Industry Location Germany*. London: Deloitte.

Schwellnus, Cyrille, Antton Haramboure, and Lea Samek. 2023a. "Policies to Strengthen the Resilience of Global Value Chains: Empirical Evidence from the COVID-19 Shock." OECD Science, Technology and Industry Policy Paper 141. Paris: Organisation for Economic Co-operation and Development.

Schwellnus, Cyrille, Antton Haramboure, Lea Samek, Ricardo Chiapin Pechansky, and Charles Cadestin. 2023b. "Global Value Chain Dependencies under the Magnifying Glass." OECD Science, Technology and Industry Policy Paper 142. Paris: Organisation for Economic Co-operation and Development.

Seneviratne, Sonia I., Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, and others. 2021. "Weather and Climate Extreme Events in a Changing Climate." In *Climate Change 2021: The Physical Science Basis*. Cambridge: Cambridge University Press.

Simchi-Levi, David. 2015. "Find the Weak Link in Your Supply Chain." *Harvard Business Review*, June 9. https://hbr.org/2015/06/find-the-weak-link-in-your-supply-chain.

Simchi-Levi, David, William Schmidt, and Yehua Wei. 2014. "From Superstorms to Factory Fires: Managing Unpredictable Supply-Chain Disruptions." *Harvard Business Review*, January-February. https://hbr.org/2014/01/from-superstorms-to-factory-fires-managing-unpredictable-supply-chain-disruptions.

Tirschwell, Peter. 2022. "Container Shipping: Supply Chains Will Remain Disrupted Well into 2022." S&P Global Market Intelligence, April 4. https://www.spglobal.com/marketintelligence/en/news-insights/research/container-shipping-supply-chains-will-remain-disrupted-well-into-2022.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5, no. 2: 207–32.

Wen, Yi. 2016. "China's Rapid Rise: From Backward Agrarian Society to Industrial Powerhouse in Just 35 Years." Blog post, April 11, Federal Reserve Bank of St. Louis Regional Economist Blog. https://www.stlouisfed.org/publications/regional-economist/april-2016/chinas-rapid-rise-from-backward-agrarian-society-to-industrial-powerhouse-in-just-35-years.

White House. 2021. *Building Resilient Supply Chains, Revitalizing American Manufacturing, and Fostering Broad-Based Growth*. Washington: White House.

White House. 2022. "Fact Sheet: Biden-Harris Administration Announces New Initiative to Improve Supply Chain Data Flow." Briefing Room Statements and Releases, March 15. Washington: Author.

World Bank. 2020. *World Development Report 2020: Trading for Development in the Age of Global Value Chains*. Washington: World Bank.

World Health Organization (WHO). 2023. "WHO Director-General's Opening Remarks at the Media Briefing—5 May 2023." https://www.who.int/news-room/speeches/item/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023.

York, Erica. 2023. "Tracking the Economic Impact of U.S. Tariffs and Retaliatory Actions." Tax Foundation, July 7. https://taxfoundation.org/research/all/federal/tariffs-trump-trade-war/.

Comments and Discussion

COMMENT BY

PINELOPI K. GOLDBERG At the Spring 2023 *Brookings Papers on Economic Activity* Conference, I presented a paper, written jointly with Tristan Reed of the World Bank, on a closely related topic, that is, deglobalization trends amid the waning political and popular support for free trade (Goldberg and Reed 2023). So unsurprisingly, this topic is close to my heart. One of the questions that we raised in that paper was "How do we increase resilience?" We suggested that, in order to make progress in answering this question, we need to start by defining what resilience is.

As a starting point, we proposed one definition that Markus Brunnermeier had put forward in his book, *The Resilient Society*: resilience is the ability to "bend but . . . not break" (Brunnermeier 2021, 2). In his book, Brunnermeier compares a reed to an oak to contrast resilience with robustness. The reed sways even with the slightest breeze but does not break when the wind is strong. In contrast, the oak can withstand light winds, but if the wind is strong enough, then it breaks. We pointed out that, for this general notion of resilience to be useful, we need to operationalize it and benchmark it. This is not something that economists can do by themselves as it involves value judgments that need to be made by the society as a whole. Finally, we emphasized that with this foundation in place, we need to figure out how to measure resilience.

Against this background, the contribution of this paper is on the measurement side. Baldwin, Freeman, and Theodorakopoulos describe an important measurement exercise. Its implications for shocks and resilience are discussed toward the end. My overall reaction is that, in order to make progress

Brookings Papers on Economic Activity, Fall 2023: 135–167

on measurement, we need to have first resolved the conceptual issues laid out above: we need to have a clear idea of what we need to measure and why.

There are many commonalities between the taxonomy presented at the end of this paper and the taxonomy presented in Goldberg and Reed (2023). For reference, I reproduce here the schematic from Goldberg and Reed (2023, 367). As we emphasize in our article, "resilience" cannot be defined without reference to a specific shock. I will not go over our taxonomy in detail given that it is explained in our *BPEA* article, but let me highlight two issues that will be important in my discussion.

Relevant considerations when defining "Resilience."

- Nature and magnitude of shock:
  - Supply, demand, or both
  - Sector-specific, country-specific, or both
  - Idiosyncratic or systemic
- Time horizon (short-, medium-, or long-run):
  - Dependent on sector (e.g., food, medicines, where time is of the essence)
  - Dependent on (possibly non-homothetic) preferences (e.g., consumers in rich countries without well-developed public transportation may consider a car a necessity)
- Level of aggregation
  - Economy
  - Industry
  - Firm
  - Household

The first is the time horizon. When we talk about resilience, are we thinking about resilience within a week, which may in fact be the appropriate time horizon if we are concerned about medical supplies? Or do we have in mind a longer time horizon, which may be more relevant if we are considering the purchase of a new car, for example?

The second issue is the appropriate level of aggregation. Is the concern about one particular plant or firm closing down in response to a shock; about a sector, a region; or about the aggregate economy?

The answers to these questions will be context-specific. At any rate, we need to have a clear idea what we are after before we attempt to measure it.

As said earlier, the contribution of this paper is to measurement. The measurement exercise is expertly done and well described in this paper as well as a companion National Bureau of Economic Research (NBER) working paper that provides many additional details (Baldwin, Freeman,

and Theodorakopoulos 2022). I will not comment on the specifics of measurement in the rest of this discussion. Instead, I will focus my comments on the implications of measurement, first for trade policy evaluation, and then for the question of resilience. But first, a brief overview of the exercise carried out in this paper.

OVERVIEW OF THE MEASUREMENT EXERCISE Briefly, what is the measurement exercise? What the paper essentially does is measuring the share of each country in intermediate input imports in the United States. There is an important distinction between the face value measure (the direct bilateral imports of intermediate goods) and the look-through measure (the imports of intermediates you get if you take into account the entire input-output structure).

For the latter, one needs inter-country input-output (ICIO) tables that provide information on the input-output relationships for the entire world. The main drawback of the input-output tables is that they are only available at the very aggregate level. The authors are clear about this limitation: in the paper, they have seventeen manufacturing sectors. The authors' main message is that, based on the look-through measure, China is much more important than one might think based on the face value measure. Not only that, but this measure has also grown. If we compare 1995 to 2018 (figure 7 in the paper), the share of China has increased substantially.

Judging this exercise in the context of the literature, one might wonder why we need yet another global value chain (GVC) measure. There is already extensive literature on measuring GVCs in trade.[1] The answer is that most of the measures (as the authors point out) were focused on measuring backward or inward integration and inferring the net value of trade. They were not focused on measuring the exposure of the domestic economy to shocks, which is the focus of the present paper.

As a side note, an interesting aspect of the earlier literature is that one of its motivations was to show that China was not as important in international trade as people thought (in gross terms, China was dominant, but less so in net value terms). In this paper, the motivation is the exact opposite, namely, to show that the dependence on China has increased substantially. It is a very different point of view.

IMPLICATIONS OF MEASUREMENT RESULTS

Implications of results for trade policy evaluation. We can debate what the paper's results mean for resilience, but a clear message is that a complete de-Chinafication of the US economy, that is, a complete decoupling

1. See, for instance, the World Bank's *World Development Report 2020* on GVCs.

of the US economy from China, may be very costly, if not impossible. Let me explain.

One of the most valuable applications of the measurement exercise is its use in the evaluation of trade policy. The recent US-China trade war provides an apt example. In 2018, the United States imposed tariffs on China, expecting a reduction of Chinese exports to the United States. This expectation was confirmed: the US tariffs and the subsequent retaliation by China reduced bilateral trade between the two countries.[2]

I have contributed to this topic myself (together with various coauthors), but a more recent paper by Alfaro and Chor (2023), presented at the Jackson Hole Symposium in August 2023, provides an up-to-date picture of the US-China trade. Their data include the latest export restrictions that the United States has imposed on China. They document that the bilateral trade between these two countries, including the bilateral imports from China to the United States, has decreased sharply. These results are interpreted as evidence that the United States' decoupling from China is happening at a fast rate.

Now, the results of the present paper cast Alfaro and Chor's (2023) results in new light. Baldwin, Freeman, and Theodorakopoulos suggest that the dependence on China may not have been reduced as much as Alfaro and Chor's face value measure suggests. As an example, consider the role of Vietnam. As Chinese imports in the United States become more expensive due to tariffs, there is a substitution in the United States toward imports from Vietnam. However, the look-through measure provided in the present paper suggests that, in order to produce these Vietnamese products, the Vietnamese need to use Chinese intermediates.

This is a different argument from the one initially made that the Chinese could simply reroute their exports through Vietnam to evade US tariffs. This is not rerouting. Instead, the point is that, to produce products in Vietnam, you need to use Chinese intermediates. In this case, an increase in the US imports from Vietnam may indirectly increase the US imports of intermediates from China.

Whether this is important or not, I do not know. I have to say that in my work with Fajgelbaum, Khandelwal, Kennedy, and Taglioni, we did not find that the Chinese global exports increased as a result of the trade war. But we did uncover some unexpected patterns regarding the response of global trade flows to the US trade war (Fajgelbaum and others 2023). One can only make sense of these patterns if one accounts for all global

2. See Fajgelbaum and others (2020, 2023).

input-output relationships and their reallocation in response to the trade war. Therefore, I believe this is an important area for research, and I am pleased to see someone working on it.

The authors are ideally positioned to address such questions. The new ICIO tables will be coming out at the end of the year. A natural next step is to repeat the exercise presented in this paper with the more recent data that reflect the recent actions of the United States vis-à-vis China and vice versa, and compute the updated look-through measures in each sector. It would be fascinating to investigate—using the look-through measures—whether the increasing US dependency on China documented in the present paper has been slowed down or reversed due to recent US trade policy. If it hasn't, this would provide support to the argument that the US trade policy has been ineffective in achieving decoupling from China, and that a de-Chinafication of the US economy may be infeasible. If it has, the natural question is how, given that the US imports from other countries (Vietnam, in my example above) use Chinese intermediates. There is no point in speculating about answers and mechanisms when we do not have the facts yet. But I am looking forward to learning about the facts based on what I hope will be the authors' next paper.

There is one caveat, however. While there is nothing the authors can do about it, it is worth keeping in mind that the sectoral level of the ICIO tables may be too aggregate to capture some of the interesting action.

The caveats associated with aggregate data in the context of GVCs are explored in De Gortari's (2019) work. De Gortari focuses on the automobile value chain of Mexico. He shows that the percentage of intermediates sourced from a particular country may be specific to the particular brand produced in Mexico and to the destination to which this brand is exported. For instance, Mexican exports of cars to the United States use on average 74 percent of US value-added. In contrast, Mexican automobile exports to Germany use only 18 percent of US value-added. So the input-output relationships are specific to each product/export destination pair.

In the present context, this means that we may not see a decline in the average share of total (direct plus indirect) intermediate input imports from China at the ICIO sectoral level, though it is conceivable that the US actions have reduced this share in more disaggregate product categories, to the extent that they have explicitly targeted such categories. I will come back to these aggregation challenges below.

Implications of results for resilience. Inferring resilience is the main motivation of the paper and an issue of great concern these days. A key figure in the paper is figure 4, which shows that the look-through share of

China in US manufacturing inputs is 3.5 percent on average, and as high as 6.3 percent in "Clothes."

Is this high or low? We do not know. This is one case where the need for a benchmark becomes apparent. This is also why I pointed out at the outset that without a benchmark in mind, it is not possible to evaluate the figures presented in this paper.

A further interpretation difficulty arises from the fact that the input shares are not sufficient statistics for dependency or resilience. They can serve as red flags. It is useful to have information on input shares, but such information is not sufficient by itself.

The issues here are analogous to those that come up these days in the discussion about competition and antitrust. Industrial organization economists have emphasized that market shares and concentration indexes are not sufficient statistics for competition. They are red flags, but one needs much more information and economic analysis to establish market power. In the present context of resilience, it is natural to associate resilience with the availability of alternatives and the ability to readily substitute toward them. But then, what one needs to judge resilience is the substitution elasticities on the demand side and the supply elasticities at a micro level. These elasticities in turn depend on the aggregation level. At a disaggregate level, many relationships and production technologies are Leontief; in contrast, substitutability could be much higher at a higher level of aggregation. Substitutability also depends on the relevant time horizon.

Coming back to my introductory remarks, this is precisely why the level of aggregation and time horizon of the analysis are important. I would argue that most of the policy issues we are concerned about these days that are related to resilience, often play out at a much more granular level than at the sectoral level of this paper's analysis. At that level, technologies are often Leontief, and average shares are not informative about the degree of dependency. Let me give three examples.

The first example comes from the work of the other discussant, Benjamin Golub (Elliott, Golub, and Leduc 2022). The authors motivate their analysis by providing a specific example of a relationship-specific investment in commercial airspace: the Airbus A380 uses a particular engine produced by Rolls-Royce, the Trent 900 engine. If Rolls-Royce has a disruption, Airbus cannot substitute, at least not in the short run, toward another engine. Someone might say, well, this is something that affects Airbus. Is that an issue that should worry all of us? In this particular case, given that the aerospace industry is an international duopoly with Airbus on one hand and Boeing on the other, it is an issue that is important, not just for Airbus and

for Europe, but also for the United States and the world as a whole. In this example, the bottleneck arising from a potential disruption plays out at a very granular level, which would not be captured in sectoral data.

The second example is from the semiconductor industry. Why is there so much concern about Taiwan? Looking at figure 4 in the paper, the US input import shares from Taiwan are below 1 percent in every sector, even when one employs the look-through measure. Based on these numbers, one would not have thought the US dependency on Taiwan to be significant. However, it turns out that about 92 percent of advanced logic capacity (i.e., semiconductor chips that are less than ten nanometers) is produced by a single company (Taiwan Semiconductor Manufacturing Corporation or TSMC) in Taiwan, while the remaining 8 percent is produced in South Korea.[3] These are the most important advanced semiconductor chips. The concern here is about a specific relationship that plays out at a very granular level.

Smartphones are another example. In a recent paper, Thun and others (2022) introduce a new concept, "massive modularity," and claim that it adequately describes the nature of many products in technologically intensive industries, such as mobile handsets (i.e., smartphones). Massive modular ecosystems (MMEs) are comprised of several interconnected functional modules that can be broken down into more specialized modules, each with its own standards, innovation potential, and market structure. While the industry as a whole is fragmented and geographically dispersed, there is extremely high market concentration at the level of each component with production being concentrated in individual countries.

This is evident in figure 9 in Thun and others (2022). The manufacturing of a mobile phone requires components from multiple regions of the world: the United States, Europe, China, Japan, South Korea, Taiwan, and others. So, at the level of the product, that is, the mobile phone, there seems to be little concentration in individual countries. But at a more granular level, the figure reveals extremely high concentration at the component or subsystem level: the market for the display component, for example, is dominated by South Korea with an 81 percent market share. On the other hand, the market for the central processing unit is dominated by the United States with a 72 percent share.

There are two key takeaways from this figure in Thun and others (2022). First, given that for any specific component there is enormous concentration, there are good reasons to be concerned about dependency and resilience.

3. See Varas and others (2021, 35).

Second, for the final product to be manufactured, one needs the cooperation of all countries involved. This makes decoupling from any specific country extremely costly.

As a sidenote, this is precisely the reason that the United States has so much power in imposing export restrictions vis-à-vis China in the semiconductor market. The United States may not be manufacturing and exporting semiconductors directly to China anymore—the manufacturing takes place in foundries located in other countries. However, the United States is still extremely important in design, software development, and specialized capital equipment used by the foundries. As a result, the United States turns out to be as important to the semiconductor global supply chains as the countries in which the foundries are located (e.g., Taiwan).

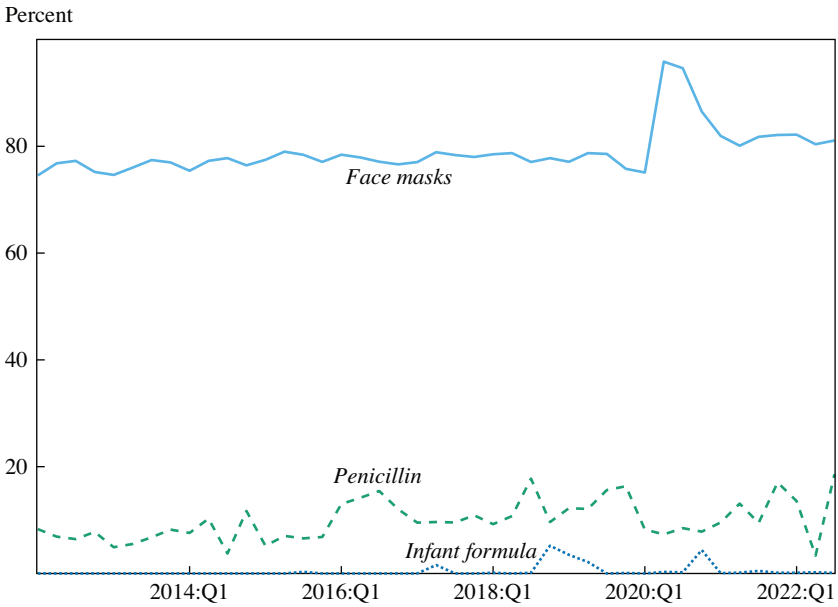
These patterns lead to the policy-related paradox eloquently described by Thun and others (2022) in the abstract of their paper: "MMEs generate strategic and geopolitical pressures for decoupling when placed under stress, but the same set of circumstances also creates pressures for maintaining the business relationships and institutions that have come to underpin global integration."

Let me now come back to a statement I made at the beginning of my discussion, namely that resilience cannot be evaluated without reference to a specific shock. Let us focus on those cases where the US dependency on China, as revealed by import shares or availability of alternative import sources, is high. As pointed out by Evenett (2020) and Goldberg and Reed (2023), such cases are rare. Figure 1 below reproduces the figure 7, panel A, of Goldberg and Reed (2023). It displays the share of US imports from non-friendly countries[4] for three critical products in the health care sector: penicillin, infant formula, and face masks. The shares of imports from non-friendly countries are minuscule for penicillin and infant formula. However, face masks follow a different pattern. Since 2012, almost 80 percent of imports of face masks come from a single non-friendly country, China. What does this imply for resilience?

The uptick in figure 1 during the second and third quarters of 2020 gives a hint at the answer. At the peak of the pandemic in the United States, imports of face masks from China increased substantially and helped alleviate domestic bottlenecks. Due to fortuitous circumstances, the first wave of COVID-19 was over in China by the time COVID-19 affected the

4. Countries are classified as non-friendly if, in the YouGov (2017) survey, less than 50 percent of Americans view the country as a friend or ally. See Goldberg and Reed (2023), notes to figure 7.

Figure 1. Percent of Imports of Medical Goods from Non-friendly Countries



Source: Reproduced from Goldberg and Reed (2023), copyright The Brookings Institution.

United States, and excess supplies of face masks in China were redirected toward the United States. This is a case where there was high dependency on China—as measured by the US import share. Nevertheless, this dependency proved beneficial during the pandemic and increased the resilience of the US economy.

Of course, the response of imports may be different in the future if we are faced with a different type of shock. But once again, the point is that resilience is not a meaningful concept without reference to the specific shock with respect to which resilience is evaluated.

The authors argue that systemic shocks are becoming more important. I am not sure what the evidence to that effect is. But even if this is the case, it is still unclear what the look-through measures imply for resilience.

I take systemic shocks in this context to mean shocks that affect multiple sectors of an economy, let's say China. If a country-specific shock hit a country as large as China, not only the United States, but the entire world would be affected. But what would such a shock plausibly be?

Broadly speaking, there are two types of shocks: natural shocks (e.g., earthquakes, tsunamis, weather-related events) that are exogenous to policy,

at least in short-term horizons; and man-made shocks, such as shocks caused by shifting geopolitics.

A natural shock is unlikely to affect a country with the geographic size of China all at once. Even COVID-19, the largest shock we have recently experienced, affected China in waves, making it more manageable and containing its international trade ramifications. On the other hand, a man-made, policy-induced shock, triggered by geopolitical tensions, is highly likely.

Given the extent of international interdependence, any action taken by China or the United States in response to a geopolitical shock would require the cooperation of multiple trade partners to be effective—it would require "weaponized interdependence," to use the term coined by Farrell and Newman (2019). If, for instance, China decided to stop supplying the US market for geopolitical reasons, then it would have to persuade other countries, such as Vietnam, to also stop exporting to the United States—otherwise Chinese exports would reach the United States indirectly via Vietnam. And vice versa, if the United States wants to be effective in containing the exports of a particular product, such as advanced semiconductor chips, to China, it needs the cooperation of all countries involved in the semiconductor global value chain (as we have seen in the past year).

Such actions would reverberate through the world trading system with potentially severe long-run effects on international trade and prosperity. But in this case, the pain would be self-inflicted in my opinion. In the presence of a high degree of international interdependence, there are two ways to increase resilience to geopolitical risk. The first is to reduce interdependence, retreating to trade among "friends." The other is to try to avoid conflict in the first place by managing, not escalating, existing tensions. Rather than rallying as many countries as possible to make trade restrictions bite, we could be encouraging international cooperation as a means to increase resilience.

CONCLUDING REMARKS To conclude, the paper offers a valuable measurement exercise that will have useful applications in the evaluation of trade policy, especially the recent actions to decouple from China. From the perspective of resilience, it is important to lay out a clear conceptual framework before attempting to assess resilience. Most importantly, the data and measures provided in this study need to be complemented by case studies of individual sectors or products that will provide a deeper understanding of the complex technologies and interdependencies at a more granular level. I hope that the present paper will inspire such work in the future.

REFERENCES FOR THE GOLDBERG COMMENT

Alfaro, Laura, and Davin Chor. 2023. "Global Supply Chains: The Looming 'Great Reallocation.'" Working Paper 31661. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31661.

Baldwin, Richard, Rebecca Freeman, and Angelos Theodorakopoulos. 2022. "Horses for Courses: Measuring Foreign Supply Chain Exposure." Working Paper 30525. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30525.

Brunnermeier, Markus K. 2021. *The Resilient Society: Economics after COVID*. Colorado Springs, Colo.: Endeavor Literary Press.

De Gortari, Alonso. 2019. "Disentangling Global Value Chains." Working Paper 25868. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w25868.

Elliott, Matthew, Benjamin Golub, and Matthew V. Leduc. 2022. "Supply Network Formation and Fragility." *American Economic Review* 112, no. 8: 2701–47.

Evenett, Simon J. 2020. "Chinese Whispers: COVID-19, Global Supply Chains in Essential Goods, and Public Policy." *Journal of International Business Policy* 3: 408–29.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. "The Return to Protectionism." *Quarterly Journal of Economics* 135, no. 1: 1–55.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, Amit K. Khandelwal, and Daria Taglioni. 2023. "The US-China Trade War and Global Reallocations." *American Economic Review: Insights*, forthcoming. (The latest working paper is available at https://www.nber.org/papers/w29562.)

Farrell, Henry, and Abraham L. Newman. 2019. "Weaponized Interdependence: How Global Economic Networks Shape State Coercion." *International Security* 44, no. 1: 42–79.

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

Thun, Eric, Daria Taglioni, Timothy Sturgeon, and Mark P. Dallas. 2022. "Massive Modularity: Understanding Industry Organization in the Digital Age—The Case of Mobile Phone Handsets." Policy Research Working Paper 10164. Washington: World Bank. https://hdl.handle.net/10986/37971.

Varas, Antonio, Raj Varadarajan, Ramiro Palma, Jimmy Goodrich, and Falan Yinug. 2021. "Strengthening the Global Semiconductor Supply Chain in an Uncertain Era." Boston Consulting Group.

World Bank. 2020. *World Development Report 2020: Trading for Development in the Age of Global Value Chains*. Washington: World Bank.

YouGov. 2017. "America's Friends and Enemies." February 2. http://today.yougov.com/topics/international/articles-reports/2017/02/02/americas-friends-and-enemies.

COMMENT BY

YANN CALVÓ LÓPEZ and BENJAMIN GOLUB The COVID-19 pandemic reminded the world of the importance of supply chains and of their fragility. From the beginning of the pandemic in early 2020 and lasting beyond the end of 2021, shortages of consumer and intermediate goods became widespread across many locations and industries. Supply chain issues have been seen as a major driver of economic volatility and inflation in the United States, the eurozone, and beyond (Helper and Soltas 2021; De Santis and Stoevsky 2023; Rubene 2023; De Guindos 2023). Baldwin, Freeman, and Theodorakopoulos (henceforth "the authors") are motivated by the challenge of understanding the structural economic factors underlying these disruptions. The authors document the exposures of US manufacturing to various industries and locales, examine the various shocks that can travel via these exposures, and discuss policy remedies.

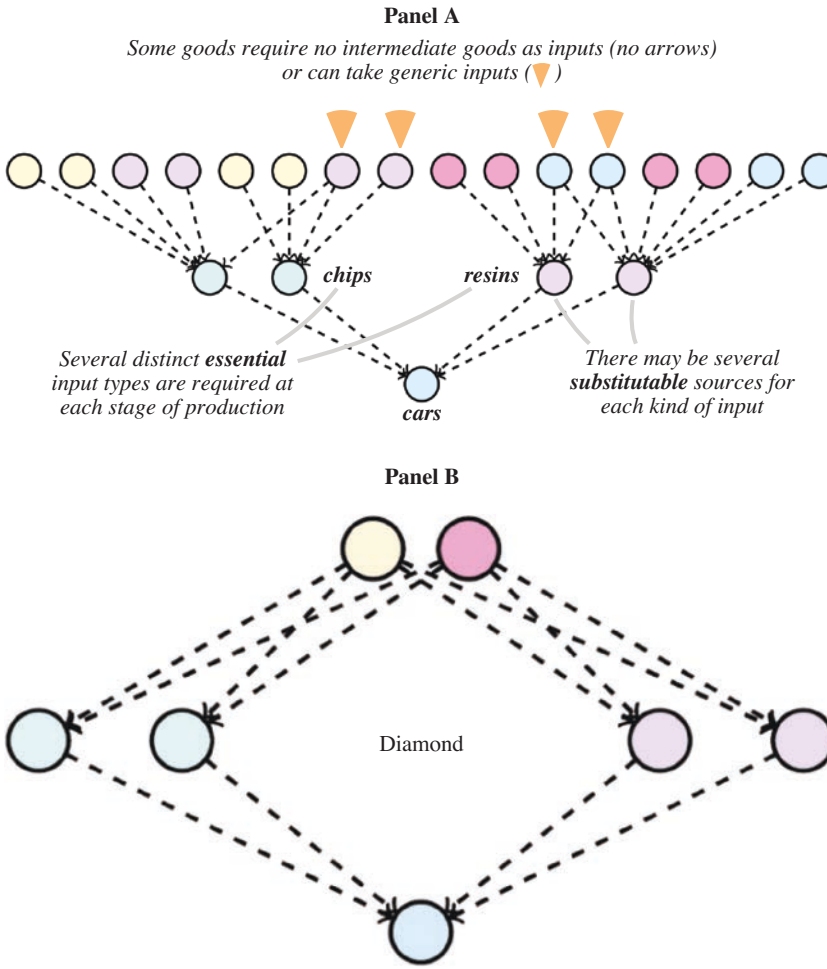
In this comment, we argue that microeconomic modeling of individual firms or plants, and their supply relationships, is essential to understanding supply chain volatility—even if the ultimate focus is macroeconomic.

To articulate this point, we first review an approach to modeling a supply network developed by Elliott, Golub, and Leduc (2022). A supply network consists of a set of firms (nodes) and sourcing relationships among them (directed links) reflecting who sources inputs from whom.[1] Input requirements can be generic or specific. Some firms can source generic inputs from a large variety of suppliers; others have customized inputs and can only get certain inputs if specific partners deliver on contracts. A firm's supply network can have many tiers—that is, a firm's suppliers may source goods from other suppliers further upstream, and so forth. In practice, looking even a few levels into such networks reveals a vast array of items and businesses, with dependencies that branch extensively—in contrast to the linear structure that is suggested by the term *supply chain*. To take a concrete example, after the Great East Japan Earthquake—a disruption whose consequences cascaded far beyond the northeast of Japan, where it started—Toyota mapped out its supply network, probing as many as ten layers of indirect dependence. This exercise uncovered 400,000 items that Toyota sources directly or indirectly (McLain 2021).[2] A schematic illustration of this kind of network is shown in figure 1, panel A.

1. In large firms, the nodes should be thought of as plants; we use the term *firms* in our exposition for simplicity.

2. Lund and others (2020) did a similar exercise for General Motors and found that it had over 17,000 indirect suppliers.

Figure 1. Firm-Level Supply Networks



Source: Authors' illustration.

Note: Panel A shows the main features of supply networks in our model: sourcing of multiple types of essential inputs by each firm (or plant); the possibility of multi-sourcing; and some nodes requiring no inputs or only generic inputs. The arrows are supply relationships. They indicate that a given firm can potentially supply an input to the firm downstream. Panel B shows an example of a diamond-shaped network. Despite the appearance of diversification in the first layer, the firm farthest downstream ultimately depends on a small group of suppliers.

The structure of a supply network describes exposures—direct and indirect—of firms to the performance of other firms. These exposures are the medium through which economic distress is transmitted from firm to firm. The ultimate source of distress is an economic shock—an exogenous disruption to some aspect of the network.

To give a sense of how such a perspective is useful, we provide some brief illustrations of supply network volatility, introducing some key aspects of both the networks and the types of disruptions they experience. Throughout this comment, we will mostly focus on discrete failures, such as a firm being unable to produce for a time, rather than a gradual degradation of performance.[3] Links may fail if relationships are disrupted—for example, by regulatory barriers to trade, failures of sourcing agreements, shipping congestion, or geopolitical conflicts. Nodes may fail when firms are temporarily unable to operate due, for example, to strikes, financing problems, or natural disasters.

Our first illustration focuses on the concentration of reliance, which occurs when a large amount of production ultimately depends on a small part of the economy—either a few firms or a specific locale. This can be seen as a diamond shape in the supply networks, as illustrated in figure 1, panel B: a firm's sourcing might look diversified through a few layers of dependence but narrows further upstream. In such a situation, regional disruptions, or even firm-specific ones, can have dramatic and distant consequences. For instance, a fire in a cleanroom at Renesas Electronics Corporation, a Japanese chip producer, contributed to a chip shortage that may have cost carmakers as much as $110 billion (Wayland 2021; Sourcengine Team 2021). Similarly, after the Great East Japan Earthquake, firms with disaster-hit suppliers experienced a 3.8 percentage point reduction in their growth rate, while firms with disaster-hit consumers experienced a 3.1 percentage point decline (Carvalho and others 2021). These results also highlight the importance of specific dependencies. Barrot and Sauvagnat (2016) find that, because of input specificity, it takes substantial time—often several months—for firms to substitute to new suppliers after idiosyncratic shocks, even when alternative sources are available. The disruptions we are interested in occur on this time scale: a supplier fails, their customers experience disruption, then that cascades to their customers, and so on.

3. The extensive literature in economics on so-called production networks, as surveyed, for example, in Carvalho and Tahbaz-Salehi (2019) and Baqaee and Rubbo (2023), typically models disruptions as continuous (i.e., sufficiently small, or at least well-modeled mathematically as being small) and uses calculus. Discrete disruptions are arguably more central to short-run supply network volatility.

Diamond-shaped dependencies are important, but they are only one of the ways that supply network structures can amplify vulnerability to shocks. Many recent supply network problems cannot be traced to cascades emanating from some salient point of failure. Baldwin, Freeman, and Theodorakopoulos offer a useful taxonomy of different kinds of shocks and then give the following sketch of the pandemic supply network crisis, highlighting a shock that is the polar opposite of an idiosyncratic shock to a firm. During the COVID-19 pandemic, there was a sudden increase in demand for consumer goods—for example, exercise machines and televisions—as consumers substituted away from in-person services to leisure at home. This spike in demand strained the global logistics system. Though it responded by shipping more goods than ever before (UNCTAD 2021), the resulting worldwide logistical issues, such as congested ports and misplaced shipping containers, had far-reaching effects. These had an impact on most shipping links, including many unrelated to the initial shock. The resulting widespread disruptions, correlated across many industries, became a central focus in the popular and business press. These disruptions constituted an aggregate shock to the links in the supply network. Our perspective is that understanding the implications of this phenomenon requires a firm-level model, combined with new insights in network theory. We will see that even well-diversified, complex networks can be very fragile in the face of aggregate shocks, starkly amplifying them (Elliott and Golub 2022), and that firm incentives can be severely misaligned with social welfare.

More broadly, we use the theoretical lens of supply networks to interrogate the facts and policy issues raised by the authors. We do this with reference to each of their main exercises: mapping exposures, modeling different kinds of shocks, and contemplating the endogenous responses of firms and policymakers. In each case, our perspective is that a model of firm-level supply networks is essential to making sense of the issues.

EXPOSURES: THE LIMITATIONS OF AGGREGATE STATISTICS The authors' main quantitative exercise is an accounting of how much various manufacturing sectors, in the United States and comparator countries, source from specific sectors in specific nations, both directly and indirectly. They primarily use input-output tables to report aggregated dependencies.[4] The discussion recounts the measurements and certain trends in them. The exercise is motivated by questions of exposure to disruptions, but the paper stops short of offering a model to make this connection precise. While we believe that the measurements are highly informative about aspects of supply networks

4. Specifically, the OECD's 2021 release of Inter-Country Input Output (ICIO) tables.

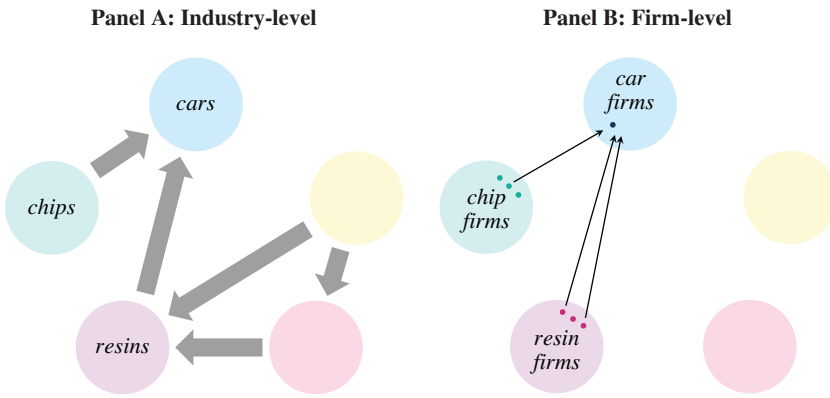
as we have defined them, they present some limitations. In this section, we interpret and critique the authors' discussion of dependencies.

Baldwin, Freeman, and Theodorakopoulos emphasize indirect exposure: for instance, an electronic component imported by the US auto industry from South Korea, constituting 15 percent of the dollar value of autos, might contain 50 percent Chinese inputs (in value terms). The paper uses the term *look-through exposure* to refer to the fraction of a sector's inputs sourced from a given industry in a given country when all indirect dependencies are accounted for. In the current example, the sourcing chain we have described would contribute 7.5 percent of US auto inputs to China. This may be contrasted with *face value exposure*, which only considers the immediate origin of intermediate inputs. Section II of the authors' paper quantifies the look-through exposures of various manufacturing sectors, revealing that these differ from, and often exceed, corresponding face value exposures. It also documents a geographic shift in look-through foreign intermediate dependencies, focusing on a concentration toward China between 1995 and 2018—the last year for which they have input-output data. More broadly, the paper contrasts the insights that can be derived from look-through exposure accounting as compared with a face value approach. It argues that the former allows for a more comprehensive picture of interdependencies than the latter.

The dynamics of exposure statistics at the industry-country level are fascinating and add much beyond the study of face value exposures. However, the dangers an economy faces due to disruptions are ultimately realized in the firm-level supply network. For this reason, our perspective is that, conceptually, the analysis must start at the disaggregated level, illustrated in figure 2, panel B. Moreover, the indirect exposures at the industry-country level are just one summary statistic of firm exposures. It is important to think through what such aggregated exposure statistics—whether face value or look-through—can and cannot tell us about how firms are affected by changes in their suppliers' functionality. In what follows, we point to several gaps between what the look-through statistics capture and what ultimately matters.

Industry-level indirect reliance can neglect across-industry substitution. The first concern with exposure accounting is that it can understate substitution possibilities, even in the short run. Across-industry substitution can play a pivotal role in avoiding catastrophic outcomes in the face of supply chain disruptions. To illustrate this, we focus on a case that Baldwin, Freeman, and Theodorakopoulos mention—that of Germany after the disruption of Russian gas supplies in the summer of 2022.

Figure 2. A Comparison of the Industry-Level versus Firm-Level Picture



Source: Authors' illustration.

Note: Panel A depicts input flows between industries. In the firm-level picture (shown in panel B), in contrast, a given firm (denoted by a small node) must use specific relationships to source from firms in other industries. Some of these links function in a given period, while others might not.

In March 2022, Russian gas accounted for around 55 percent of Germany's gas consumption. Citing reports that Germany was profoundly dependent on Russian gas, the German government did not sever ties with Russia following the start of the Russian invasion of Ukraine. Nonetheless, by the end of the summer of 2022, Germany stopped receiving Russian gas when Gazprom, the main Russian state-owned gas company, discontinued its supply. Surprisingly, Germany only experienced a "technical mini-recession" during the subsequent winter (Moll, Schularick, and Zachmann 2023, abstract). This outcome sharply diverged from some earlier forecasts, which had predicted a 6 to 12 percent drop in Germany's GDP in the event of a total embargo on Russian gas (Moll, Schularick, and Zachmann 2023).

In addition to some alternate sourcing (e.g., increased imports of liquefied natural gas), input substitution across energy sources was crucial in mitigating the impact of a shock of this type, as extensively documented by Moll, Schularick, and Zachmann (2023). The point here is a familiar one—that input-output tables are just a snapshot. The exposures documented might reflect rigid technological constraints that create a severe dependence, but they also might be easily bypassed when needed. In fact, there turned out to be firms that were already set up to source energy without Russian gas; these firms had the capacity to expand production, and orders could shift to them. These aspects of firm-level production structure were essential to Germany's surprising resilience.

Value-weighted exposure mapping understates firm-level vulnerability. While an industry-level exposure snapshot can understate substitution possibilities and the resilience of an economy, it can also understate important rigidities. As we have already mentioned, customization is a big part of how firms get their parts, and firms often fail to quickly find alternative suppliers when it is necessary (Barrot and Sauvagnat 2016). Moreover, as the just-cited paper emphasizes (building on a large body of literature), modern production involves strong complementarities in inputs: a missing part disables the productive use of many others.

These facts together imply that if a firm is missing a low-cost, low-value-added item, such as certain cheap microchips, major disruptions can ensue (Elliott and Jackson 2023). Such an item, however, would barely show up in the exposure statistics since these statistics are value-weighted at market prices. From the macroeconomic perspective, a cheap good cannot stop high-value production. But this perspective misses rigidities that are central to volatility on the timescale of several quarters. The fact that a firm can find another supplier of a disrupted input at a low price in three months does not render it operational now.[5]

Dangerous foreign reliance, or beneficial diversification? Behind the descriptive statistics in section II of the paper, an issue of seemingly obvious policy interest is the increased exposure of the United States and several similar economies to imports. As the authors note, whether such exposure is good or bad is unclear. We elaborate on this point here and put it in the context of our supply network perspective.

Let us focus, for concreteness, on the issue of US (direct and indirect) exposure to Chinese inputs. While "country" is a natural unit for accounting purposes, it is not clear that concentrated sourcing at the country level is concentrated in the ways that ultimately matter. Sourcing from a large country could potentially provide considerable robustness. In particular, conditional on sourcing many inputs from China, the extent of geographical concentration within China matters. If sourcing is narrowly focused on specific areas, then US production can be exposed to highly localized shocks. On the other hand, if sourcing is diversified within China, that could provide

5. Baldwin, Freeman, and Theodorakopoulos recognize the importance of disaggregating in studying exposures at the product level in section II.F. This analysis, however, is limited by available data. They use detailed export and import statistics published by the US Census Bureau, but these have two important limitations: they do not contain information on which sector imports the goods and do not distinguish between intermediate and final goods. Moreover, such data are informative only about face value exposure—they only consider the direct source of intermediate inputs.

considerable protection against idiosyncratic risk, though not against distinctively political or otherwise nationally correlated risks.

The takeaway is that the decision to carry out exposure mapping at a specific level, such as the country level, should be supported by an explicit account of why we worry about exposure at that particular level or at least why that level offers a reasonable proxy for the issue of interest.

A summary. The unifying message of this section is that look-through exposures should be seen as a summary statistic of a complex microeconomic reality underneath—that of the firm connections. Despite their usefulness in depicting possible sources of supply chain fragility, they offer only a partial accounting of many important features of supply networks. In the remainder of this comment, we discuss how exposure mapping can be used in conjunction with shock modeling to understand some salient supply network risks.

SHOCKS: THE SOURCES OF DISRUPTION To analyze how reliance shapes resilience and to design interventions, we must model the shocks or potential disruptions the network faces. Baldwin, Freeman, and Theodorakopoulos develop a very useful typology of supply chain shocks. Here, we review it and then discuss a particular aspect of it that we think deserves deeper theoretical and empirical study.

The authors classify shocks into three different sources:

- *Supply shocks* refer to events or situations that cause significant disruptions or disturbances in the availability or production of goods and services within a supply chain.
- *Demand shocks* refer to sudden and significant changes in demand for products and services that affect the supply chain.
- *Connectivity shocks* refer to significant disruptions or disturbances in the interconnected and interdependent networks that facilitate the movement of inputs within the supply chain.

They cross this classification with a division of shocks into two types:

- *Idiosyncratic*: These are firm-specific or otherwise highly localized disruptions that affect one supply chain, as opposed to broader, market-wide disturbances. They are typically unforeseen and can arise from internal or external factors specific to the firm's operations, relationships, or environment.
- *Systemic*: Systemic shocks are large-scale disruptions that affect multiple companies, industries, or even entire economies. These shocks are characterized by their widespread impact across the global supply chain network.

ZOOMING IN ON CONNECTIVITY Connectivity, from the first axis of the taxonomy, seems especially important to understanding the 2020–2022 shortages, as well as supply chain volatility more generally. Nevertheless, we see this concept as understudied relative to its importance.

Connectivity encompasses much more than just logistical links. Let us dig down into several dimensions of connectivity and the economic factors that determine it. The first dimension consists of technological relationships. The large-scale structure of the supply network depicted in figure 2, panel B, is shaped both by technological facts and by firms' choices of which of many possible "recipes" to use in producing goods (Boehm and Oberfield 2020). For example, a clothing manufacturer can have workers sew buttons onto clothing by hand or buy specialized machines for this purpose. Firms' choices here, in turn, are influenced by things like what kind of software is available to help them plan and integrate production across firms, and whether standards exist that help harmonize production processes. Another choice is multi-sourcing: how many alternative (potential) suppliers does a firm have access to for a certain input? A closely related but softer part of connectivity concerns relational contracts. In the face of potential disruptions, which can be very costly (Hendricks and Singhal 2003, 2005a, 2005b), firms invest in relationships. These investments include favors such as ordering in advance to assist a supplier during a period of low demand (Uzzi 1997) and the allocation of scarce supply to a customer in need (Carlton 1978). They also include a variety of noncontractible activities to stabilize and facilitate relationships; an important outcome of these activities is building interpersonal trust. Legal and contractual frameworks also play a significant role. They form a base for connectivity. Finally, there is the logistics and shipping aspect of connectivity, which is the most familiar: the systems and services that move goods from one place to another. These interact in obvious ways with the previous aspects.

Connectivity shocks correspondingly include a range of disruptions. An idiosyncratic shock to relational connectivity might consist of a contract breaking down due to debt nonpayment. Idiosyncratic logistical shocks include fires and misplaced shipping pallets.[6] On a broader scale, Brexit is an example of an aggregate shock to both relational contracts and the logistics network. Increased bureaucracy and changes in rules and regulations have made it difficult for many UK firms to deal with their EU counterparts (British Chambers of Commerce 2021). Similarly, an aggregate logistical

6. Hendricks and Singhal (2003, 2005a, 2005b) show that localized disruptions are often associated with durable declines in sales growth and stock returns.

shock can manifest as congestion at points of entry such as tunnels or ports, leading to delayed deliveries for many industries at once (Murray 2023; Komaromi, Cerdeiro, and Liu 2022).[7]

A conceptual challenge. The discussion above makes clear that one type of shock can lead to another. Demand shocks can lead to connectivity shocks. For instance, the demand shock during the COVID-19 pandemic led to a connectivity shock (port congestion, etc.). These shocks, in turn, seemed to seriously affect aggregate supply, motivating the theory of Elliott, Golub, and Leduc (2022). Including such effects in models is clearly important. However, such issues have not received much attention in standard macroeconomic models, and this presents an important challenge for researchers. Indeed, standard models do not even have a standard abstraction for capturing the object to which connectivity shocks happen. We might call this object *connectivity capital*. An adequate notion of connectivity capital should ultimately be rich enough to include the various dimensions discussed above.

It is worth remarking on the reason that we call connectivity a type of capital. We do this because many of its dimensions can be seen as produced factors of production that are not fully depleted in the course of particular production processes.[8]

RESPONSES TO SHOCKS: FIRM BEHAVIOR AND PUBLIC POLICY The consequences of shocks are a concern for firms as well as for policymakers at the subnational, national, and international levels. Both types of actors make many choices that affect both the structure of firm supply networks and the probability of shocks occurring. Their choices thus shape the robustness of the economy.

Firms' incentives in making these choices may be misaligned with the social interest in aggregate robustness. Indeed, Baldwin, Freeman, and Theodorakopoulos sketch some theoretical ideas concerning why the incentives of firms to mitigate risks might not be aligned with those of a social

7. Technological compatibility is rarely shocked in the short run, but in the longer run, advances in information technology, such as AutoCAD modeling and enterprise resource planning systems, have reshaped how firms interact.

8. Connectivity also relies on a variety of services and human capital inputs. It is tempting to take a minimal approach and incorporate connectivity as simply a complement to shipping services. At a minimum, this would have to be done in a modern production network model (Baqaee and Farhi 2019, 2020), since in the old-school models, Hulten's theorem applies and the quantitative estimates of the harm of negative shipping shocks seem severely understated (because shipping value added at usual prices is low). But beyond this, connectivity shocks can be amplified in distinctive ways—an issue studied by Elliott, Golub, and Leduc (2022) and Acemoglu and Tahbaz-Salehi (2023).

planner.[9] They argue that firms might invest less in robustness than is socially optimal because they are less risk-averse than a planner. Our view is that this perspective is insufficiently precise for understanding the issues distinctive to supply chain risk. The basic premise is not even generally true: a social planner is often much less risk-averse over the fortunes of any given firm than individual firm decision-makers, because small firms make only a small relative contribution to aggregate outcomes. What is true is that social planners are more risk-averse over disasters where many firms fail at once, or where supply is severely disrupted. But then what is key is whether firms fail in a correlated way, and understanding that requires more detailed modeling.

The supply network perspective provides an organizing framework. To make this point, we focus particularly on connectivity shocks, though the analysis extends to other types of shocks. Misalignment of incentives arises in all of the various chosen aspects of connectivity we have emphasized above—firms' choices of inputs and multi-sourcing, as well as their management of relational contracts and logistics. We now analyze these misalignments, bringing the above-discussed typology of shocks together with a firm-level approach to exposure mapping.

Decisions about suppliers. Perhaps the most fundamental connectivity decisions made in the economy are firms' sourcing decisions. These have large consequences from the standpoint of robustness. For example, if a firm ends up having high indirect dependence on a single region, it might end up highly vulnerable to regional supply or logistics shocks.

Firms' incentives in making these decisions need not be aligned with those of a planner. For example, in choosing their suppliers, many firms might prefer to source from a single region because of economies of scale and scope in setting up sourcing relationships. Moreover, and probably more importantly, the lowest-cost suppliers, with the highest short-run productivity, might all be located in one region, for example, to benefit from agglomeration externalities (Duranton and Puga 2004; Rosenthal and Strange 2004). Even in the absence of colocation of a firm's immediate suppliers, a more dispersed set of suppliers might rely on the same upstream providers (as in the diamond-shaped network example discussed earlier). In either case, a single regional shock could simultaneously disrupt many

9. We use the construct, familiar in economic theory, of a fictitious entity—the social planner—that makes decisions aimed at maximizing some notion of social surplus. This construct is helpful for understanding distortions that cause individual decisions to differ from what such a planner would do.

firms that have arranged their sourcing this way, resulting in widespread fragility across the supply network.

The key tension between individual and social interests is that the planner is concerned with the correlation of firms' performance, whereas each individual firm is concerned only with its own performance and profitability. Whether this is a problem or not depends on whether firms' sourcing incentives push their performance to become highly correlated.

How much to invest in a given link's robustness. Beyond choosing whom to link with, firms invest in making links with their suppliers more robust and resilient. They might, for instance, invest in their logistics departments —for instance, by using technologies to monitor shipments and communicate about disruptions. They can also store more inventory (so as to compensate for temporary disruptions by having extra inputs on hand).[10] Finally, they can undertake investments in their relationships by optimizing both relational and formal contracts.

Such investments protect firms against shocks to the performance of their relationships. In other words, these investments are especially suited to safeguard firms against connectivity shocks. However, as Elliott, Golub, and Leduc (2022) show, there are circumstances in which firms have too little incentive to invest in relationship strength, compared to what is socially optimal.

To make this point, Elliott, Golub, and Leduc (2022) work with a version of the supply network model sketched earlier in this comment. In the model, each firm can invest in robustness and thereby improve its relationship strengths, defined as the probability that each relationship will be functional in a given time period. They give conditions under which it is optimal for firms to invest less in robustness than what would be socially optimal. This leads to inefficient supply chain vulnerabilities: the economy has a substantial probability of ending up in a configuration where small, systemic shocks affecting the functioning of supply relationships have stark, amplifying effects.[11] A planner controlling link investments, on the

10. The management of inventory has been an important concern in the field of operations. Running a "just-in-time" strategy with low inventories reduces costs (Callen, Fader, and Krinsky 2000). Keeping more inventory allows firms to weather logistical shocks better. But when a firm sources a large number of complex inputs, customized to evolving production, managing risk through inventory can become impractical (Goodman and Chokshi 2021).

11. A key condition for this result to hold is the widespread customization of intermediate inputs or, in other words, a lack of short-run substitution. As previously mentioned, there is good evidence that firms do indeed struggle to substitute for new suppliers in the timescale of one or two quarters (Barrot and Sauvagnat 2016).

other hand, would never choose to make the economy vulnerable to such fragility.

Summing up. A reliable instinct of academic economists is to imagine a certain fictitious complete-markets benchmark in order to illuminate what missing market is preventing the efficient allocation of resources. In our setting, the complete-market benchmark would entail the existence of securities allowing bets on every conceivable event (e.g., every possible pattern of shocks), along with some additional assumptions, for example, that the mathematical descriptions of firms' production possibilities are sufficiently well-behaved. In such a paradise, market equilibria would exist in which all risk would be correctly priced, and social interests in firms' reliability could be transmitted to them via the price mechanism.

Such markets do not and probably could not exist due to the sheer vastness of vagueness of the space of possible shocks. It is a natural theoretical question whether markets that are somewhat more realistic could mitigate incentive misalignment. For example, could incentives be improved by dynamic markets where firms that survive are allowed to gouge their customers to some extent? We are not optimistic that this would offer a robust solution.[12]

What is clear is that the investments firms endogenously make toward robustness generally differ from what is socially optimal. A firm-level analysis is important for revealing both this divergence and the factors driving it. And within that type of analysis, we argue that connectivity capital and shocks to it are likely to play an outsized yet understudied role. In the next section, we make one more argument for that position, using a policy issue that motivates Baldwin, Freeman, and Theodorakopoulos.

WHY FEAR EXPOSURE TO CHINA? Baldwin, Freeman, and Theodorakopoulos are clearly interested in exposure to countries—with China playing a particularly central role due to its rise as an important indirect supplier. We have emphasized that the right network to focus on is at the firm level. And we have also noted that, at this level, it is not obvious why country-level exposures are especially significant. For instance, a large country such as China might offer unusually good opportunities for multi-sourcing and, for US firms, additionally provide insurance against domestic shocks.

It seems clear that concern over reliance on Chinese inputs must stem from the anticipation of country-level shocks to commercial relationships that Chinese firms have with their counterparties. Such shocks could arise from tariffs or geopolitical and military tensions. However, even once we

12. See Elliott and Golub (2022) for a fuller discussion.

focus on such shocks, it still needs to be explained why US economic policymakers should be especially worried about the extent of *indirect* exposure to China. After all, it seems implausible that China would, or could, prevent the use of any of its inputs indirectly in US goods. For example, Russian energy remains an input into a great deal of production by countries sanctioning Russia after its 2022 full-scale invasion of Ukraine, while Russia indirectly buys many goods made in the European Union and the United States—including ones that are banned from directly buying.

The perspective of connectivity capital introduced above can nevertheless help rationalize concerns about exposure to China. The example of Brexit helps motivate the point. Brexit disrupted trade relations and the workings of commerce—by increasing regulatory hurdles, for example. The resulting effects have been widely discussed as a damper on European and UK trade and economic performance.[13] While the US relationship with China is much more arm's-length than the pre-Brexit relationship between Europe and the United Kingdom, increasing tension with China could have similar adverse consequences, degrading the performance of many links, including those between China and various non-US economies that supply the United States. Systemic damage to commerce within Asia and across the Pacific would be one of the main ways a China-related crisis would have an impact on supply networks.

The most natural way to view this is as a connectivity shock to many supply networks. We have discussed above the distinctive and severe ways in which these can be amplified. Properly describing these connectivity shocks in economic models and explaining why and how we should be concerned about them (beyond the rough sketch we have given) requires further developing our understanding, both theoretical and empirical, of supply networks. What is clear is that documenting growing indirect exposure is just a first step.

CONCLUDING DISCUSSION Our main message is that modeling of supply networks at the firm level is indispensable to understanding supply-chain volatility, even when the overarching focus is macroeconomic. Most of the interesting questions about supply chains and indirect exposures cannot be usefully analyzed while staying at a highly aggregated level.

We started by reviewing the authors' exposure mapping, discussing both its usefulness and aspects of exposure that are not captured by it—ones that require a firm-level analysis. We then reviewed and extended their

13. Office for Budget Responsibility, "Brexit Analysis," https://obr.uk/forecasts-in-depth/the-economy-forecast/brexit-analysis/.

typology of supply chain shocks, emphasizing the need for proper modeling of connectivity capital—the (multidimensional) object that is degraded when connectivity shocks happen. Next, we turned to a discussion of misalignments between firms and a social planner in incentives to invest in connectivity. Finally, we circled back to a focal policy concern of Baldwin, Freeman, and Theodorakopoulos: the dependence of the United States on Chinese intermediate inputs. We argued that the perspective of supply networks and their connectivity shocks is critical to making sense of why this may merit concern.

Broadly, the authors make clear the importance of supply network issues in understanding current economic trends. We have argued that these issues raise an urgent need for better concepts and theories of firm-level sourcing relationships and their disruptions. This poses an important challenge at the intersection of network theory and macroeconomics, which we hope will prove energizing to researchers.

REFERENCES FOR THE CALVÓ LÓPEZ AND GOLUB COMMENT

Acemoglu, Daron, and Alireza Tahbaz-Salehi. 2023. "The Macroeconomics of Supply Chain Disruptions." Working Paper. https://economics.mit.edu/sites/default/files/2023-10/The%20Macroeconomics%20of%20Supply%20Chain%20Disruptions.pdf.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87, no. 4: 1155–203.

Baqaee, David Rezza, and Emmanuel Farhi. 2020. "Productivity and Misallocation in General Equilibrium." *Quarterly Journal of Economics* 135, no. 1: 105–63.

Baqaee, David Rezza, and Elisa Rubbo. 2023. "Micro Propagation and Macro Aggregation." *Annual Review of Economics* 15: 91–123.

Barrot, Jean-Noël, and Julien Sauvagnat. 2016. "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks." *Quarterly Journal of Economics* 131, no. 3: 1543–92.

Boehm, Johannes, and Ezra Oberfield. 2020. "Misallocation in the Market for Inputs: Enforcement and the Organization of Production." *Quarterly Journal of Economics* 135, no. 4: 2007–58.

British Chambers of Commerce. 2021. "Almost Half of Firms Facing Difficulties Trading with EU under Post-Brexit Trade Agreement." December 23. https://www.britishchambers.org.uk/news/2021/12/almost-half-of-firms-facing-difficulties-trading-with-eu-under-post-brexit-trade-agreement.

Callen, Jeffrey L., Chris Fader, and Itzhak Krinsky. 2000. "Just-in-Time: A Cross-Sectional Plant Analysis." *International Journal of Production Economics* 63, no. 3: 277–301.

Carlton, Dennis W. 1978. "Market Behavior with Demand Uncertainty and Price Inflexibility." *American Economic Review* 68, no. 4: 571–87.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi. 2021. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." *Quarterly Journal of Economics* 136, no. 2: 1255–321.

Carvalho, Vasco M., and Alireza Tahbaz-Salehi. 2019. "Production Networks: A Primer." *Annual Review of Economics* 11: 635–63.

De Guindos, Luis. 2023. "The Inflation Outlook and Monetary Policy in the Euro Area." Keynote speech at King's College London, July 7. https://www.ecb.europa.eu/press/key/date/2023/html/ecb.sp230707~8f8f9debc6.en.html.

De Santis, Roberto A., and Grigor Stoevsky. 2023. "The Role of Supply and Demand in the Post-Pandemic Recovery in the Euro Area." *ECB Economic Bulletin* 4. https://www.ecb.europa.eu/pub/economic-bulletin/articles/2023/html/ecb.ebart202304_01~509fc9d72c.en.html.

Duranton, Gilles, and Diego Puga. 2004. "Micro-Foundations of Urban Agglomeration Economies." In *Handbook of Regional and Urban Economics, Volume 4*, edited by J. Vernon Henderson and Jacques-François Thisse. Amsterdam: Elsevier.

Elliott, Matthew, and Benjamin Golub. 2022. "Networks and Economic Fragility." *Annual Review of Economics* 14: 665–96.

Elliott, Matthew, Benjamin Golub, and Matthew V. Leduc. 2022. "Supply Network Formation and Fragility." *American Economic Review* 112, no. 8: 2701–47.

Elliott, Matthew, and Matthew O. Jackson. 2023. "Supply Chain Disruptions, the Structure of Production Networks, and the Impact of Globalization." Working Paper. Social Science Research Network, October 27. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4580819.

Goodman, Peter S., and Niraj Chokshi. 2021. "How the World Ran Out of Everything." *New York Times*, June 1. https://www.nytimes.com/2021/06/01/business/coronavirus-global-shortages.html.

Helper, Susan, and Evan Soltas. 2021. "Why the Pandemic Has Disrupted Supply Chains." Blog post, June 17, The White House, Council of Economic Advisers. https://www.whitehouse.gov/cea/written-materials/2021/06/17/why-the-pandemic-has-disrupted-supply-chains/.

Hendricks, Kevin B., and Vinod R. Singhal. 2003. "The Effect of Supply Chain Glitches on Shareholder Wealth." *Journal of Operations Management* 21, no. 5: 501–22.

Hendricks, Kevin B., and Vinod R. Singhal. 2005a. "Association between Supply Chain Glitches and Operating Performance." *Management Science* 51, no. 5: 695–711.

Hendricks, Kevin B., and Vinod R. Singhal. 2005b. "An Empirical Analysis of the Effect of Supply Chain Disruptions on Long-Run Stock Price Performance and Equity Risk of the Firm." *Production and Operations Management* 14, no. 1: 35–52.

Komaromi, Andras, Diego A. Cerdeiro, and Yang Liu. 2022. "Supply Chains and Port Congestion around the World." Working Paper 2022/059. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/03/25/Supply-Chains-and-Port-Congestion-Around-the-World-515673.

Lund, Susan, James Manyika, Jonathan Woetzel, Edward Barriball, Mekala Krishnan, Knut Alicke, and others. 2020. *Risk, Resilience, and Rebalancing in Global Value Chains*. New York: McKinsey Global Institute.

McLain, Sean. 2021. "Auto Makers Retreat from 50 Years of 'Just in Time' Manufacturing." *Wall Street Journal*, May 3. https://www.wsj.com/articles/auto-makers-retreat-from-50-years-of-just-in-time-manufacturing-11620051251.

Moll, Benjamin, Moritz Schularick, and Georg Zachmann. 2023. "The Power of Substitution: The Great German Gas Debate in Retrospect." *Brookings Papers on Economic Activity (BPEA)* Conference Draft, Fall. https://www.brookings.edu/wp-content/uploads/2023/09/6_Moll-et-al_unembargoed_updated.pdf. (The final version is included in the present volume of *BPEA*.)

Murray, Brendan. 2023. "Brexit Costs, Delays Still Weigh on UK Companies Trying to Trade." Bloomberg, January 10. https://www.bloomberg.com/news/newsletters/2023-01-10/supply-chain-latest-uk-companies-still-suffer-from-brexit-delays.

Rosenthal, Stuart S., and William C. Strange. 2004. "Evidence on the Nature and Sources of Agglomeration Economies." In *Handbook of Regional and Urban Economics, Volume 4*, edited by J. Vernon Henderson and Jacques-François Thisse. Amsterdam: Elsevier.

Rubene, Ieva. 2023. "Indicators for Producer Price Pressures in Consumer Goods Inflation." *ECB Economic Bulletin* 3. https://www.ecb.europa.eu/pub/economic-bulletin/focus/2023/html/ecb.ebbox202303_05~e6d2439ece.en.html.

Sourcengine Team. 2021. "Automobile Industry Estimated Costs from Global Chip Shortage Revised up to $110B." May 25. https://www.sourcengine.com/blog/automobile-industry-estimated-costs-from-global-chip-shortage-revised-up-to-110b.

United Nations Conference on Trade and Development (UNCTAD). 2021. "Shipping during COVID-19: Why Container Freight Rates Have Surged." April 23. https://unctad.org/news/shipping-during-covid-19-why-container-freight-rates-have-surged.

Uzzi, Brian. 1997. "Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness." *Administrative Science Quarterly* 42, no. 1: 35–67.

Wayland, Michael. 2021. "Chip Shortage Expected to Cost Auto Industry $110 Billion in Revenue in 2021." CNBC, May 14. https://www.cnbc.com/2021/05/14/chip-shortage-expected-to-cost-auto-industry-110-billion-in-2021.html.

GENERAL DISCUSSION Şebnem Kalemli-Özcan emphasized the importance of timing in understanding macroeconomic dynamics, providing the example that goods that are considered substitutable in the long

run may not be substitutable over the course of a few quarters, ultimately having an impact on macroeconomic aggregates such as inflation. She asked whether economists should be utilizing more micro-level to macro-level analysis to better understand macroeconomic indicators and dynamics like inflation and unemployment.

Georg Zachmann inquired whether the heterogeneity of companies that are users of inputs should be more strongly considered when measuring the impact of political shocks on the aggregate economy. Zachmann explained that companies can vary greatly in their productivity for a given input, and a shock to this input could result in the loss of further production from low value-added producers. He noted that this would decrease input consumption but leave a significant share of the value-added unaffected, creating a buffer against a supply crisis.

Elaine Buckberg emphasized that the private sector is not monolithic in its supply chain management strategies, such as multiple sourcing or inventory management, which play an important role in maintaining a competitive advantage. Buckberg stressed the importance of considering the duration of shocks, highlighting that the ability to endure a shock is more important than just its source.

Rebecca Freeman agreed that a distinction between the duration of shocks versus their source needs to be made. She noted that her coauthors and she tried to address this by creating a distinction between resilience and robustness—where resilience is the speed of recovery after a crisis, while robustness is related to where, and in which areas, failure is not acceptable.

Angelos Theodorakopoulos touched on the discussions of time horizon and heterogeneity by bringing up Pinelopi Goldberg's discussant remarks regarding the need to define resilience and exposure metrics, in addition to benchmarks, before making progress on measurement. Theodorakopoulos also drew attention to comments about large firms' relationships with their suppliers, pointing out that on the aggregate level, countries and industries are dependent upon a whole network of supply that should also be considered when thinking about trade exposure.

Tarek Hassan drew attention to the challenges of shock modeling and the potential value of quarterly data from firm executives' reports to financial markets. Hassan noted that by analyzing earnings calls, he was able to monitor the impact of business supply shocks and the propagation through final goods manufacturers, consumer durables, and industrials during the COVID-19 pandemic.

David Romer expressed his confusion regarding gross trade flow measures. He presented a hypothetical scenario where authors exchanged

hundreds of versions of a paper via email across international borders, with each version differing only trivially from the previous one, and where a simple analysis based on gross trade flows would lead to the obviously incorrect conclusion that these flows increased the paper's value by several hundredfold. He argued that having clear objectives in designing trade metrics is important for work in this field, though he noted he was unsure of what exactly these measures might look like.

John Haltiwanger raised concerns about the timeliness of the input-output (IO) data, noting that Bureau of Economic Analysis had released comprehensive revisions to GDP accounts that day (September 28, 2023).[1] He stated that this was the first time the 2017 Economic Census had been used for the GDP accounts, noting that the reference year was 2012 until these revisions. Haltiwanger expressed concern about the data being too old and wondered if there were potential solutions or improvements to address the timeliness problem in the data.

Freeman agreed with points made about the coarseness and timeliness of the IO data, pointing out that the international organizations and academic institutions responsible for curating the data need to do lots of time-consuming preprocessing, on top of the fact that countries report data at different points in time. Freeman highlighted that one of the main advantages of the measure that they propose in their work is to create a more macroscopic view of differences between face value trade and look-through exposure. She noted that the IO data also allowed the authors to circumvent two important caveats regarding data at the product level used for analyses of the US economy and automotive sector. Freeman explained that the product-level data do not include which sector is importing a given good nor whether the good is an intermediate or final good—which are critical pieces of information for analyses. Freeman remarked that the trends they observed have been steady over time, which can serve as a benchmark when considering some of the timeliness issues.

Richard Baldwin responded to concerns about some of the caveats surrounding IO analyses, such as the lack of substitution in a Leontief production function, recognizing that these are important considerations to account for when interpreting the authors' findings. Baldwin noted that a single measure, such as Leontief inverse, does not necessarily summarize

1. Bureau of Economic Analysis, "Gross Domestic Product (Third Estimate), Corporate Profits (Revised Estimate), Second Quarter 2023 and Comprehensive Update," https://www.bea.gov/news/2023/gross-domestic-product-third-estimate-corporate-profits-revised-estimate-second-quarter.

all matrix information. He continued, mentioning that it would be worth analyzing the exact shape of trade networks to gain insights into fragilities and single points of failure. Baldwin also stated that he believed a computable general equilibrium would be necessary to provide advancements on the authors' analysis, particularly when allowing for substitutability between geographic origins and between products. He cautioned, however, that a model of this size can become too complex to intuitively understand. Baldwin noted that he thinks of their measures as a first-order approximation that can identify areas of dependencies worth further investigating for risk, while cautioning against directly interpreting dependencies as risk.

Romer asked whether an analysis focusing on market failures would in fact lead to the conclusion that an unregulated market results in insufficient resilience. He stated that while there is of course pervasive imperfect competition, he did not see this as obviously leading to an economy that is systematically less resilient than is socially optimal. He presented the example of a company like Airbus being concerned about preserving monopoly rents, which could lead to greater supply chain resiliency relative to what a social planner would choose. Romer concluded that before policymakers potentially intervene to increase resilience, there is a need for greater attention to the relevant basic microeconomic theory.

Jason Furman wondered whether the government taking an interest in resilience can lead to perverse incentives and greater risk-taking from private firms as a result. Furman emphasized the importance of understanding the cases in which the government interest in resilience will move firms down the risk-reward curve, to less supply chain risk exposure, in addition to the cases in which intervention will possibly have the opposite effect.

Wendy Edelberg presented her working hypothesis on the need for policy interventions to enhance firms' resilience. First, Edelberg noted that the emergence of relatively new and increasing risks from geopolitical and climate-related factors necessitates additional resilience. Second, she drew attention to managerial incentives, pointing out that during good times, managers tend to underinsure their firms because they are penalized for performing worse than their peers. Edelberg highlighted that during aggregate shocks, underinsuring is also incentivized as managers are not particularly penalized for poor performance, further supporting the need for additional policy measures.

Henry Aaron brought up inventories as a way to manage risk, stating that they should be possible to implement in large swaths of the economy, despite potentially being costly. Aaron pointed out that the private sector's calculations about the value of inventories may underestimate their social

value, presenting a possible basis for government intervention to encourage more inventory holding. He stressed the need for empirical research to better understand the costs associated with carrying inventories across the manufacturing and industrial spectrum.

Baldwin thanked discussant Benjamin Golub for drawing attention to the fundamental source of shocks and noted that differentiating between supply, demand, and connectivity shocks can be useful when tailoring policy responses to different shocks. Baldwin provided the example of stockholding, stating that the policy is resilient to all three kinds of shocks. He noted that various countries have adopted stockholding in some form, citing the Strategic Petroleum Reserve and the Swiss government's subsidization of retailer stockholding in named products. Baldwin contrasted stockholding with geo-diversification, which he explained only works in supply shocks and not demand shocks. He highlighted that greater domestic production to reduce risk may even have the opposite effect, depending on the shock source. Baldwin agreed with Golub, stating that assessing a policy's cost and benefits before action is essential.

Iván Werning stated that if geopolitical risk is at the core of this work, it would be interesting to perform an analysis from the perspective of China to determine their supply chain dependence and resilience. Werning drew attention to the difference between mutual and one-sided trade dependence, noting that this could change the thinking about US-China trade relations. Hoyt Bleakley contrasted the authors' findings with Mainland China a generation ago, hypothesizing that the direct and look-through exposure measures would have been close to zero then. To him, this suggested that the long-term elasticity of substitution might be high, which would mean long-term policies like de-Chinafication—that is, policies reducing US dependence on China—could be easier to implement.

Martin Baily commented that he believed that a large degree of supply chain difficulties during the COVID-19 pandemic were due to a significant shift in demand from services to goods, recognizing that the production issues in China also played an important role. Baily said previously he had thought that China's low value-added exports were not a major concern because the value-added was lower than the gross trade. He had reconsidered this, given the authors' analysis of look-through exposure, stating that China's assembly power could grant them significant influence as they are the last producer of a finished item. Baily also said the authors' work made him reconsider the value of single-supplier models, such as vertical keiretsu, that form close relationships with suppliers to improve quality and productivity. Baily noted that while it might be acceptable to maintain close

single-supplier relationships for domestic supply, the benefits of multiple suppliers may outweigh the drawbacks when considering supply shocks and trade stability.

Robert Gordon drew attention to the rapid rise in China's prominence as a producer of finished and intermediate manufactured goods as well as the near zero growth in US manufacturing productivity over roughly the past decade.[2] Gordon expressed that he did not see a connection between Chinese intermediate imports and the lack of US productivity. He stated that, like Baily, he would have expected Chinese imports to be skewed toward lower value-added products, thus replacing US firms that produced low-value goods. This loss in low-productivity firms should theoretically have led to higher productivity in manufacturing, the absence of which puzzled Gordon.

Freeman touched on the asymmetric role of China in global supply chains, highlighting a companion paper in which they found that all major manufacturing countries are highly dependent on China—sourcing at least 2 percent of their total domestic and foreign inputs from China.[3] Freeman pointed out that China's role has declined because, although it has built up its industrial bases, becoming a major world supplier of industrial inputs, it is increasingly sourcing those inputs in its own economy domestically.

2. US Bureau of Labor Statistics, "Manufacturing Sector: Real Sectoral Output for All Workers [PRS30006041]," retrieved from FRED.

3. Richard Baldwin, Rebecca Freeman, and Angelos Theodorakopoulos, "Horses for Courses: Measuring Foreign Supply Chain Exposure," working paper 30525 (Cambridge, Mass.: National Bureau of Economic Research, 2022), https://www.nber.org/papers/w30525.

ŞEBNEM KALEMLI-ÖZCAN
University of Maryland

FILIZ UNSAL
Organisation for Economic Co-operation and Development

Global Transmission of Fed Hikes: The Role of Policy Credibility and Balance Sheets

ABSTRACT Contrary to historical episodes, the 2022–2023 tightening of US monetary policy has not yet triggered financial crisis in emerging markets. Why is this time different? To answer this question, we analyze the current situation through the lens of historical evidence. In emerging markets, the financial channel–based transmission of US policy historically led to more adverse outcomes compared to advanced economies, where the trade channel fails to smooth out these negative effects. When the Federal Reserve increases interest rates, global investors tend to shed risky assets in response to the tightening global financial conditions, affecting emerging markets more severely due to their lower credit ratings and higher risk profiles. This time around, the escape from emerging market assets and the increase in risk spreads have been limited. We document that the historical experience of higher risk spreads and capital outflows can be largely explained by the lack of credible monetary policies and dollar-denominated debt. The improvement in monetary policy frameworks combined with reduced levels of dollar-denominated debt have helped emerging markets weather the recent Federal Reserve hikes.

Conflict of Interest Disclosure: Şebnem Kalemli-Özcan holds unpaid advisory positions at the Federal Reserve Bank of New York and the Bank for International Settlements. The authors did not receive financial support from any firm or person for this paper or from any firm or person with a financial or political interest in this paper. Other than the aforementioned, the authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper. The Organisation for Economic Co-operation and Development (OECD), Filiz Unsal's employer, had the right to review this work prior to publication. This paper should not be reported as representing the official views of the OECD or of its member countries.

Brookings Papers on Economic Activity, Fall 2023: 169–225

> Contrary to many analysts' expectations, emerging markets have not spiraled into a debt crisis. This can be partly attributed to central banks' decision to reject populist policy proposals in favor of a modern iteration of macroeconomic orthodoxy.
>
> —Ken Rogoff, "The Stunning Resilience of Emerging Markets"

In stark contrast to the 1980s and 1990s, emerging markets have demonstrated resilience in the face of monetary policy tightening in advanced economies, notably the United States, during the post-COVID-19 era. Historically, sharp increases in policy rates in the United States have led to falling currencies elsewhere combined with capital outflows—the so-called sudden stops—which often resulted in widespread financial stress and crises in emerging markets and developing economies. The 1982–1983 debt crisis in Latin America, following the Federal Reserve hikes during disinflation under Paul Volcker, remains the classic example, but there are also other instances such as the 1994 tightening of US monetary policy paving the way to Asian crisis and the infamous taper tantrum of 2013. However, the recent tightening cycle has unfolded differently. This time, the majority of emerging markets have effectively navigated the most significant tightening in the United States in several decades without much damage to their economies.

What explains this newfound resilience to the US monetary policy shocks? We argue that the resilience of emerging markets comes largely from their improved monetary policy credibility, combined with a reduction in dollar borrowing. Monetary policy credibility and debt denominated in foreign currencies (FX), mostly dollars, are domestic vulnerabilities that are often linked. Weak private and public sector balance sheets due to the dollar debt and local currency assets can force central banks to defend the currency to avoid local currency depreciations, which would otherwise increase the debt burden and defaults.[1] An inflation-targeting central bank can lose its credibility by responding to exchange rate fluctuations through policy rates without a clear framework, since such behavior could entail a deviation from the "do what you say, say what you do" rule that captures the essence

1. Since most of the foreign currency debt in emerging markets and developing economies is in US dollars, reducing the extent of foreign currency debt means they borrow less in dollars relative to the 1980s and 1990s (McCauley, McGuire, and Sushko 2015).

of monetary policy credibility.[2] Our new credibility index quantifies these types of deviations within an existing framework, where most of the frameworks are centered on inflation targeting. Thus, credibility is measured through transparency, coherency, and consistency among policy tools and objectives.

While the benefits of central bank independence and inflation-targeting frameworks have been extensively highlighted in the literature using cross-country data, it is rare to quantify the improvements in policy credibility for a given country over time. We use a brand-new data set based on a narrative approach from Unsal, Papageorgiou, and Garbers (2022) to quantify the monetary policy frameworks, and hence the credibility improvements in countries over time that are exogenous to both the US monetary policy shocks and other domestic policy changes within countries. This data set is hand-collected from thousands of central bank legal documents from fifty countries over 2007–2021, to characterize the monetary policymaking across three pillars of independence and accountability: policy, operational strategy, and communications. Even though the changes in domestic monetary policy rate could be endogenous to US monetary policy and other policy and institutional changes in the country, our measure is orthogonal to such changes since it is designed to capture policy design and implementation features that enable and guide the conduct of monetary policy, rather than specific endogenous monetary policy actions at any point in time.[3]

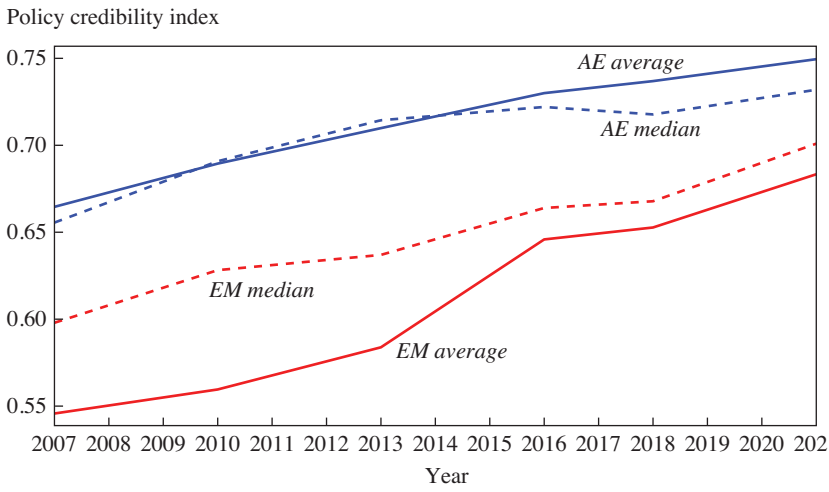
Empirical literature on the central bank independence focuses on the political independence by constructing cross-country measures and relating them to inflation and inflation expectations.[4] The theoretical underpinning of

2. There could be reasons to intervene in the exchange rate market. Our point is that, if not done correctly with a clear framework, monetary policy credibility could be jeopardized. An increasing number of emerging markets have moved toward approaches where multiple tools are employed in pursuit of multiple objectives related to financial stability, exchange rate stability, and capital flow management. See Basu and others (2020) on how an "integrated" approach helps provide macroeconomic and financial stability in the face of risk-off shocks.

3. The policy credibility index goes far beyond classifying countries' monetary or exchange rate regimes. For example, in addition to checking whether a country has a numerical target (on inflation) or not, the assessment metric considers whether the numerical target is a viable nominal anchor by encapsulating various key elements such as how the target is set and by who, the time horizon, and whether objectives and the numerical target in communications are consistent with the ones in policy and operational strategy. See the table in online appendix A.1 for an illustration of how transparency, coherence, and consistency principles underpin our credibility metric, using the criteria on the numerical targets of monetary policy as an example.

4. See, for example, Alesina and Summers (1993) and Dincer and Eichengreen (2014).

Figure 1. Policy Credibility over Time



Source: IAPOC index from Unsal, Papageorgiou, and Garbers (2022)

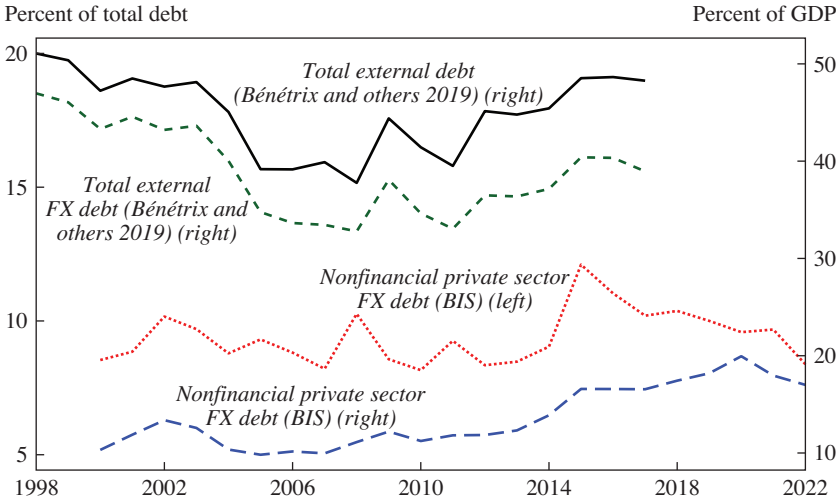
Note: The measure of policy credibility, on a scale of zero to one, is based on the monetary policy frameworks index (IAPOC index) from Unsal, Papageorgiou, and Garbers (2022). The graph shows the average and median policy credibility in advanced economies (AEs) and emerging markets (EMs) from 2007 to 2021.

this idea that delegating monetary policy to an independent body mitigates the inflationary bias comes from Rogoff (1985). Separately, there is a strand of literature starting with the work of Sargent and Wallace (1981) that studies structural models of monetary-fiscal interactions. In this line of work, fiscal dominance is interpreted as low monetary policy credibility since politicians can get central banks to finance deficits through inflation. However, there remains a gap in both theoretical and empirical literature regarding how improvements in monetary policy credibility affect emerging markets over time, especially when they face external shocks with considerable impact on their exchange rates, such as the changes in US monetary policy.

The new credibility index is plotted in figure 1. The index is between zero and one, where a value of one indicates perfect credibility. It reveals that the monetary policy credibility substantially improved in emerging markets, for both the average and median countries. In contrast, advanced countries, which already had high monetary policy credibility in 2007, showed only minimal improvement over time.

This advancement in credibility among emerging markets is paralleled by a decrease in dollar-denominated debt. Figure 2 plots the ratio of total external debt to gross domestic product (GDP) and the ratio of total external

Figure 2. Foreign Currency Debt in Emerging Markets



	FX debt/total debt 1995–2002, private sector
Argentina	0.61
Brazil	0.28
Chile	0.30
Colombia	0.25
Hungary	0.21
Mexico	0.60
Peru	0.50
Turkey	0.55

Source: Bank for International Settlements (nonfinancial private sector debt); and Bénétrix and others (2019) (total external debt and total external FX debt).

Note: Credit in US dollars to the nonfinancial private sector is estimated as the total credit in US dollars minus international debt securities for government and financial institutions, normalized by total debt and by annual GDP. We plot the averages for emerging markets and use a balanced panel in each series. The table shows the data from Di Giovanni and others (2022), Salomao and Varela (2022), Kamil (2004), Kalemli-Özcan, Kamil, and Villegas-Sanchez (2016), Aguiar (2005), and Kalemli-Özcan (2022), which are all based on confidential data from each central bank, as reported in these papers.

debt in FX to GDP. These series show some decline at first, from around 50 percent to 38 percent of GDP between 1998 and 2008, but both increased afterward during the quantitative easing in advanced economies following the global financial crisis that drove capital flows to emerging markets. As explained above, historically, what triggered central banks in emerging markets to defend their currencies in the face of Fed hikes was the FX debt-related vulnerabilities in their nonfinancial private sectors. Hence, we also

plot in figure 2 the FX debt of the nonfinancial private sector (household and corporate) both as a percentage of GDP and as a percentage of total debt. Unfortunately, the time series for these data is only available after 2000. What is remarkable is that the nonfinancial sector FX debt is below 20 percent of GDP and around 10 percent of total debt. This is a huge reduction given the historical values before the 2000s as shown in the table. There are some countries such as Turkey and Argentina, where the shares of corporate sector FX debt are still similar to the historical values, hovering around 50 percent of GDP or total debt (Di Giovanni and others 2022; Das and others 2020). But these countries would be outliers rather than the norm as of 2020. We do not analyze the FX debt of financial institutions since this debt is hedged by several regulatory restrictions. By now these ensure the FX mismatches on bank and financial intermediary balance sheets are fully hedged or minimal (IMF 2022).

There is extensive literature on the international transmission of US monetary policy, starting with Diaz-Alejandro (1983) and Calvo, Leiderman, and Reinhart (1993, 1996) that emphasize the impact of interest rate differentials between a given country and the United States on the demand for government bonds.[5] Consistent with this early literature's focus on the interest rate differentials, more recent literature on the US monetary policy spillovers to other countries has shifted attention to the financial channel of US policy transmission—switching demand of assets between the United States and the rest of the world—from the trade channel—switching demand for goods produced in the United States to those produced in the rest of the world (Rey 2013; Kalemli-Özcan 2019; Degasperi, Hong, and Ricco 2023; Chari, Dilts Stedman, and Lundblad 2021; Di Giovanni and Rogers 2023).

A prevailing finding in this body of research is the link between the changes in US monetary policy and the cross-border correlations of macro-financial conditions, that is, the global financial cycle proxied by global-level risk indicators, like the CBOE Volatility Index (VIX), the broad US dollar index, and the US excess bond premium (Bekaert, Hoerova, and Duca 2013; Rey 2013; Miranda-Agrippino and Rey 2020; Bruno and Shin 2015; Obstfeld and Zhou 2022). Hence, the underlying factors for the financial transmission channel of US monetary policy are changes in risk-taking incentives and the associated risk premia. Central to this discussion is the role of time-varying deviations from the uncovered interest parity (UIP)—

5. See also Eichengreen and Portes (1987), Reinhart and Reinhart (2009), and Reinhart and Rogoff (2009).

the country-level risk premia priced by international investors—which has been identified as crucial in understanding the deteriorating macro conditions in emerging markets with risk-sensitive capital flows (Kalemli-Özcan 2019; Di Giovanni and others 2022).[6] Based on this empirical literature, the recent theoretical works focusing on the optimal policies for emerging markets single out the UIP wedge as the key factor to be stabilized to maximize welfare (Basu and others 2020; Bianchi and Lorenzoni 2022; Itskhoki and Mukhin 2022).

The financial channel is more pronounced in distinguishing the impact of US monetary policy tightening on advanced economies versus emerging markets. This is primarily due to global investors moving away from risky assets in response to tighter global financial conditions. Emerging markets, typically considered riskier investments in any portfolio, are particularly affected by this shift. This risk-based channel underscores the significance of domestic vulnerabilities in emerging markets. We argue that the literature on the international transmission of US monetary policy overlooked a key domestic vulnerability, that is, the role of monetary policy credibility, while focusing solely on the exchange rate or the monetary policy regime. The choice of the exchange rate regime is endogenous to policy credibility: countries lacking monetary policy credibility often opt to peg their currency to the US dollar as an alternative nominal anchor. In addition, since the late 1990s, most emerging markets have moved away from pegged exchange rate regimes. Comparing countries with fixed versus floating regimes over time will identify the impact of US monetary policy on a select set of countries suffering from a time-varying selection bias.[7]

There are other variables that are likely to be endogenous to improved monetary policy credibility such as capital flows, UIP premia, inflation, exchange rates, and current accounts. We also investigate these outcomes, recognizing that many of them depend on the presence of dollar-denominated

6. See also quantitative models, where exogenous UIP deviations take center stage, such as Dedola, Rivolta, and Stracca (2017) and Akinci and Queralto (2023); see Gourinchas (2018) on the contractionary effects of US monetary policy on real outcomes of other countries. Kalemli-Özcan and Varela (2021) investigate the empirical determinants of endogenous UIP deviations, and Akinci, Kalemli-Özcan, and Queralto (2021) model such deviations in a global general equilibrium framework.

7. Dedola, Rivolta, and Stracca (2017) point out that one reason why they do not find a strong role for exchange rate regimes in driving the international spillovers of US monetary policy shocks is that none of the countries in their sample has been in a peg all the time. Iacoviello and Navarro (2019) also find exchange rate regimes inconsequential when considering higher US interest rates on economic activity.

debt. Therefore, our analysis differentiates countries not only by their monetary policy credibility, but also by their levels of dollar-denominated debt, following Kalemli-Özcan (2019).

Our broad analysis covers fifty-nine countries using quarterly data from 1990:Q1 to 2019:Q4. We analyze the recent 2021–2023 period separately. We show that, historically, the worse effects of the Fed hikes such as declining GDP, depreciating exchange rates, higher risk spreads, and higher UIP premia combined with capital outflows, can be explained by lower monetary policy credibility and higher levels of FX debt in the corporate sector.[8] We show that the improvement in these two key domestic vulnerabilities has led to a minimal impact of the Fed hikes on emerging markets so far.

The paper is composed of five sections. Section I lays out the broader literature and shows descriptive evidence. Section II details the data. Section III undertakes an empirical analysis that shows the heterogeneous effects of US monetary policy. Section IV analyzes the recent post-pandemic inflation episode and the effects of Fed hikes during this period. Section V concludes.

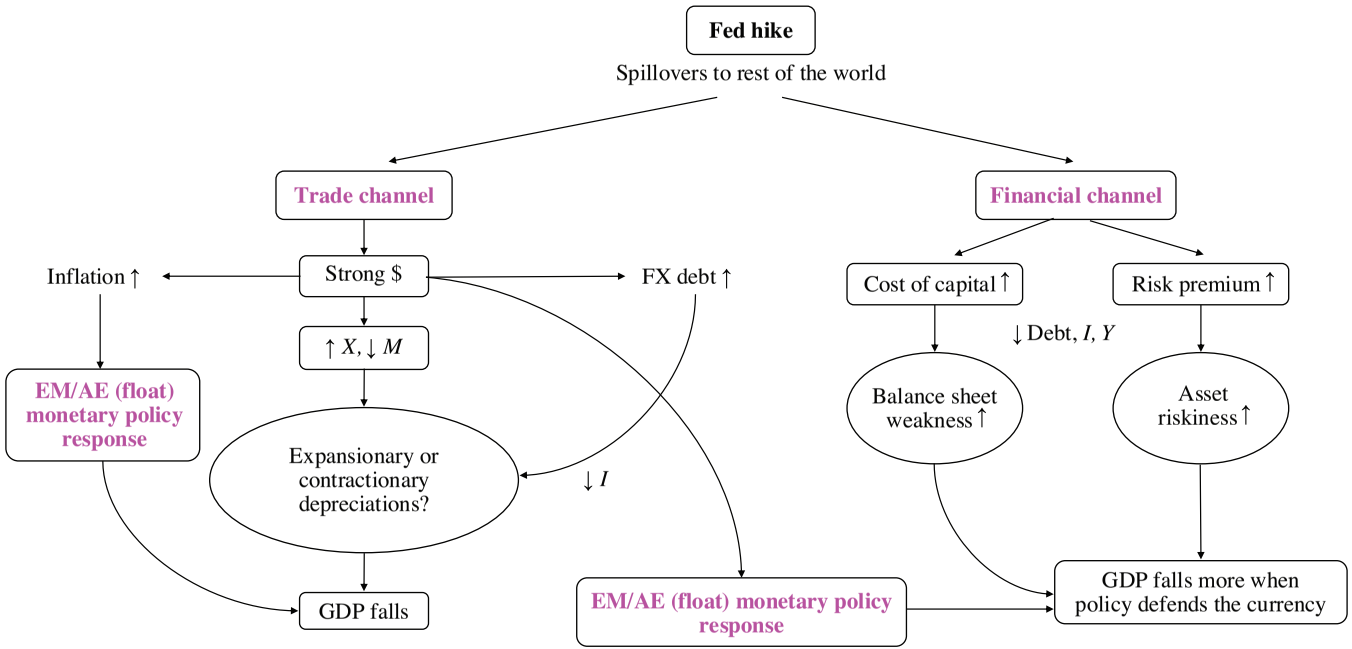
I. The Narrative within the Broader Literature

For the transmission of US monetary policy, trade and finance linkages represent two critical channels that have garnered significant attention among academics and policymakers. Figure 3 illustrates these channels and the way the literature evolved in trying to understand these channels both theoretically and empirically.

In the traditional models and empirical work, the focus was on the currency depreciations of other countries vis-à-vis dollar appreciations, akin to the Mundell-Fleming model. A currency depreciation has the potential to stimulate net exports, creating an expansionary effect, but it can also trigger inflation through exchange rate pass-through (Burstein and Gopinath 2014; Forbes, Hjortsoe, and Nenova 2018), potentially requiring monetary tightening that might lead to a contraction. When the Federal Reserve hikes the federal funds rate and the US dollar appreciates, the demand for goods switches from the now expensive US goods to the goods from the rest of the world, which suffer from a local currency depreciation but can enjoy an increase in output thanks to higher net exports. Existing evidence on this issue goes against the notion of an expansionary effect when countries' currencies depreciate and capital flows out during Fed hikes.

8. Kalemli-Özcan (2019) shows similar results for the detrimental effects of US monetary policy and risk-off shocks in high FX debt countries.

Figure 3. Fed Hike



Source: Authors' elaboration.

Figure 3 shows this as the trade channel, depicted on the left side of the diagram. The failure to find an expansionary effect of currency depreciations has been justified by the models and evidence showing the dollar pricing of exports (Gopinath 2016) or negative balance sheet effects due to currency mismatch involving unhedged dollar debt and local currency assets (Krugman 1999; Schneider and Tornell 2004; Aghion, Bacchetta, and Banerjee 2001; Cook 2004; Céspedes, Chang, and Velasco 2004; Aguiar 2005; Kalemli-Özcan, Kamil, and Villegas-Sanchez 2016). Even though there is an increase in net exports as capital flows out on net, such expenditure switching fails to initiate an expansion in output, leading to a contraction in GDP (Mendoza and Yue 2012; Gopinath and Neiman 2014) via lower investment. Consistently, Miranda-Agrippino and Rey (2020) and Obstfeld (2015) argue that the flexible exchange rates fail to fully absorb external shocks through expenditure switching. Hence, even though the trade channel is not responsible for the worse outcomes in emerging markets (falling output and capital outflows) resulting from Fed hikes, it is not smoothing out these effects either.[9]

Currency mismatches in balance sheets have often pushed policymakers to defend the currency (Calvo and Reinhart 2002; Reinhart 2000; IMF 2022) by mimicking the Fed hikes, which might intensify the contraction in their own economies. Kalemli-Özcan (2019) shows that countries that hike the policy rate to defend their currencies experience deeper recessions.

The financial channel is depicted on the right side of figure 3. The US interest rate increase not only results in higher safe rates globally, increasing the cost of capital, but also leads to higher risk premia toward inherently riskier assets such as emerging markets. As the balance sheets of US/global financial intermediaries weaken (Gertler and Kiyotaki 2010) with the Fed hikes—recently witnessed during the banking stress of 2023 (Jiang and others 2023)—they may not want to bear more risk by being exposed to

9. At the same time, countries with fixed exchange rate regimes are shown to be more sensitive to global risk shocks and a strong dollar due to higher US interest rates rather than flexible regimes, so flexible exchange rates must be doing some smoothing (Obstfeld and Zhou 2022). Kalemli-Özcan (2019) shows that this smoothing is from risk-absorbing properties of the floating exchange rates. Since the exchange rate depreciates, vis-à-vis the US dollar, the risk premia, measured as the UIP premia, on emerging market assets do not have to go up as much, limiting capital outflows and contractionary effects. Similarly, Fukui, Nakamura, and Steinsson (2023) show that exchange rate depreciations can be expansionary, not due to expenditure switching linked to higher net exports, but rather through the financial channel, when the country experiences a boom financed with capital inflows, implying a lower UIP premium.

emerging market assets, which are likely to depreciate. Thus, global investors want to dump risky assets, given higher risk aversion and a risk-off sentiment, inducing risk premia shocks for emerging markets combined with dollar appreciations.[10] As a result, asset riskiness and balance sheet weakness can go hand in hand in limiting international financial intermediation (Gabaix and Maggiori 2015).

As discussed in the earlier literature starting with the work of Diaz-Alejandro (1983), capital flows are central to both channels in the context of Fed hikes. Any resiliency to these hikes has to come from the fact that, when the Federal Reserve hikes the interest rates, emerging markets do not experience sudden stops or capital outflows; and if they do, resilience means that the extent is much smaller such that it does not affect their domestic economies. During the 1980s and 1990s, the main form of borrowing by other countries involved their sovereigns issuing dollar bonds. As shown by Alfaro, Kalemli-Özcan, and Volosovych (2014) and Kalemli-Özcan (2019), since the early 2000s, there has been a compositional change from sovereign to private sector borrowing in emerging markets, while many developing economies still rely heavily on sovereign borrowing, which dominates their capital flows (Avdjiev and others 2022). Also, the currency of borrowing has evolved, as shown by Du and Schreger (2016) and Hofmann, Patel, and Wu (2022), such that the emerging market sovereigns are increasingly borrowing in local currency, whereas the private sector, especially the nonfinancial corporations, can still only access foreign funding in US dollars as they cannot issue bonds in local currency, unlike their governments.[11] Thus, the transmission mechanism of US monetary policy might also have changed, as private capital flows are generally more sensitive to the global risk aversion. Forbes and Warnock (2012) study the total gross flows as the sum of private sector and government borrowing, and show the increasing importance of global risk factors after the mid-1990s. Avdjiev and others (2019, 2022) show that this risk sensitivity in gross flows is driven by private capital flows.

10. See models formalizing this financial channel endogenously in Jiang, Krishnamurthy, and Lustig (2021), Bianchi, Bigio, and Engel (2021), Akinci, Kalemli-Özcan, and Queralto (2021), and Devereux, Engel, and Wu (2023). Gourinchas and Rey (2022) model this story as a rise in risk aversion, and Kekre and Lenel (2021) as flight to safety.

11. These changes may indicate the shift of "original sin" from sovereigns to corporations—a term referring to the inability to issue external debt in domestic currency, coined by Eichengreen and Hausmann (1999) and Eichengreen, Hausmann, and Panizza (2005).

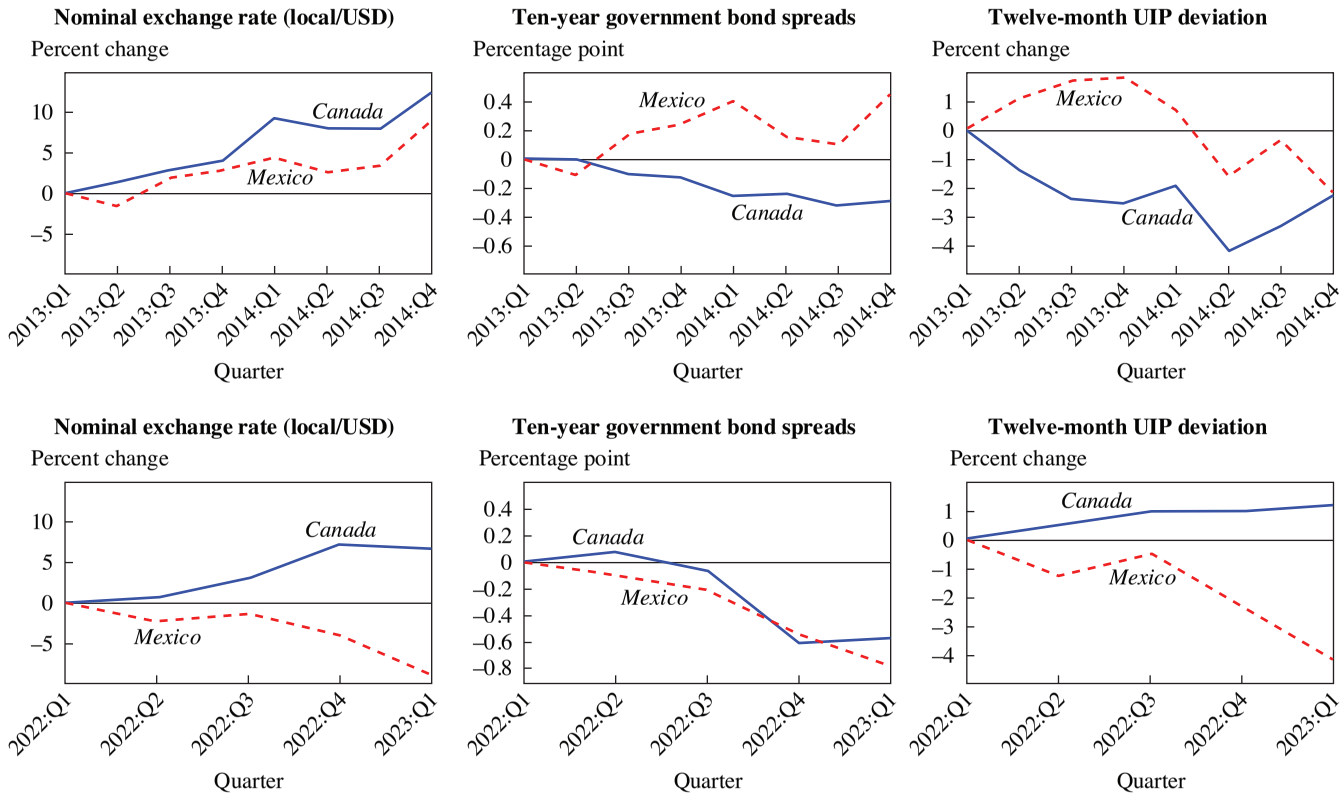
I.A. A Tale of Two Countries: Mexico and Canada

To illustrate, we use the two trading partners of the United States, Canada and Mexico, as case studies. These are both small open economies with important differences relevant to our analysis. From the perspective of the trade channel for US monetary policy transmission, the distinction between Mexico and Canada is less important; however, from the perspective of the financial channel, failing to distinguish between a small open economy and an emerging market/developing economy is detrimental.

Figure 4 documents a specific US monetary policy tightening episode, known as the taper tantrum, in May 2013, during which the Federal Reserve signaled the end of quantitative easing and an anticipated earlier increase in interest rates. Mexico and Canada, both neighboring the United States under a trade agreement, should observe a similar impact through the trade channel given both of their currencies depreciate vis-à-vis the US dollar: the nominal exchange rate depreciations, shown for Mexico and Canada, are similar. However, the risk spreads show stark contrast. During this period, the long-term risk premium in Mexico experienced a sharp increase and remained elevated for a prolonged period, captured by the ten-year government bond spreads. The short-term risk premium also rose sharply, captured by the twelve-month UIP premium. Both spreads remained mainly flat for Canada, with a slight decrease in the UIP premium. Notice that the long-term government bond spreads can capture the dollar premium via default risk if issued in dollars, or the term premium if issued in local currency. The short-term UIP premium captures the local currency premium, that is, the excess currency returns due to currency risk. The UIP premium is measured in logs as follows: $(i_{mex/can} - i_{US}) - (\Delta E(s))$, where the interest rate differential term between Mexico/Canada and the United States uses the twelve-month government bond rates in local currency, and the second term is the expected change in the peso/dollar (or Canadian dollar to US dollar) exchange rate (s) in the next twelve months.

The increase in the UIP premium for Mexico can be driven by three different channels: (1) an expected appreciation captured by a fall in the second term, $\Delta E(s)$, as currency depreciated on impact with the Federal Reserve's actions; (2) an increase in the interest rate differential above and beyond the movements in the expected exchange rate, driven by the possible response of the Mexican central bank hiking its own interest rates more than the Federal Reserve to defend the currency; or (3) a higher risk premium reflected in the interest rate differential demanded by global investors of risky Mexican assets. Kalemli-Özcan (2019), Kalemli-Özcan and Varela

Figure 4. Canada and Mexico after the Fed Hikes: Taper Tantrum versus COVID-19



Source: IMF International Financial Statistics; Bloomberg; Consensus Economics; and authors' calculations.

Note: The top row shows the evolution of variables relative to the taper tantrum (2013:Q1). The bottom row shows the evolution of variables relative to the recent Fed hikes (2022:Q1).

(2021), and De Leo, Gopinath, and Kalemli-Özcan (2022) show that it is the third channel that drives the higher UIP premium in emerging markets as a response to the US monetary policy shocks and risk-off shocks.[12]

As shown in figure 4, for 2022:Q1–2023:Q1, the recent experiences of Canada and Mexico are very different from the earlier episode. Now both countries behave in a similar way in terms of risk spreads. The Mexican exchange rate appreciated during the recent Fed hikes, implying an expected depreciation in the future. Hence, the UIP premium fell in Mexico more than in Canada, implying a lower risk premium for Mexico by global investors to hold on to the Mexican assets. The long-term risk spreads fell for both countries.[13]

I.B. A Tale of Won and Weakened Credibility: The Case of Turkey

Next, we conduct a within-country analysis to understand the changes of monetary policy credibility over time and how this could relate to macroeconomic performance, with a specific focus on Turkey. Figures 5 and 6 plot the key macro variables together with inflation dynamics, risk spreads, and changes in our policy credibility measure. Turkey serves as an effective case study for understanding the exogeneity of our policy credibility measure and its time series changes being orthogonal to the domestic and US policy changes.

After the triple crises in 2001 (balance of payments, sovereign, and banking), Turkey successfully moved to a floating exchange rate regime within an inflation-targeting framework. This framework had been in place since 2002 and during the entire period we look at; however, the implementation of inflation targeting is what drives the time variation in our credibility measure.

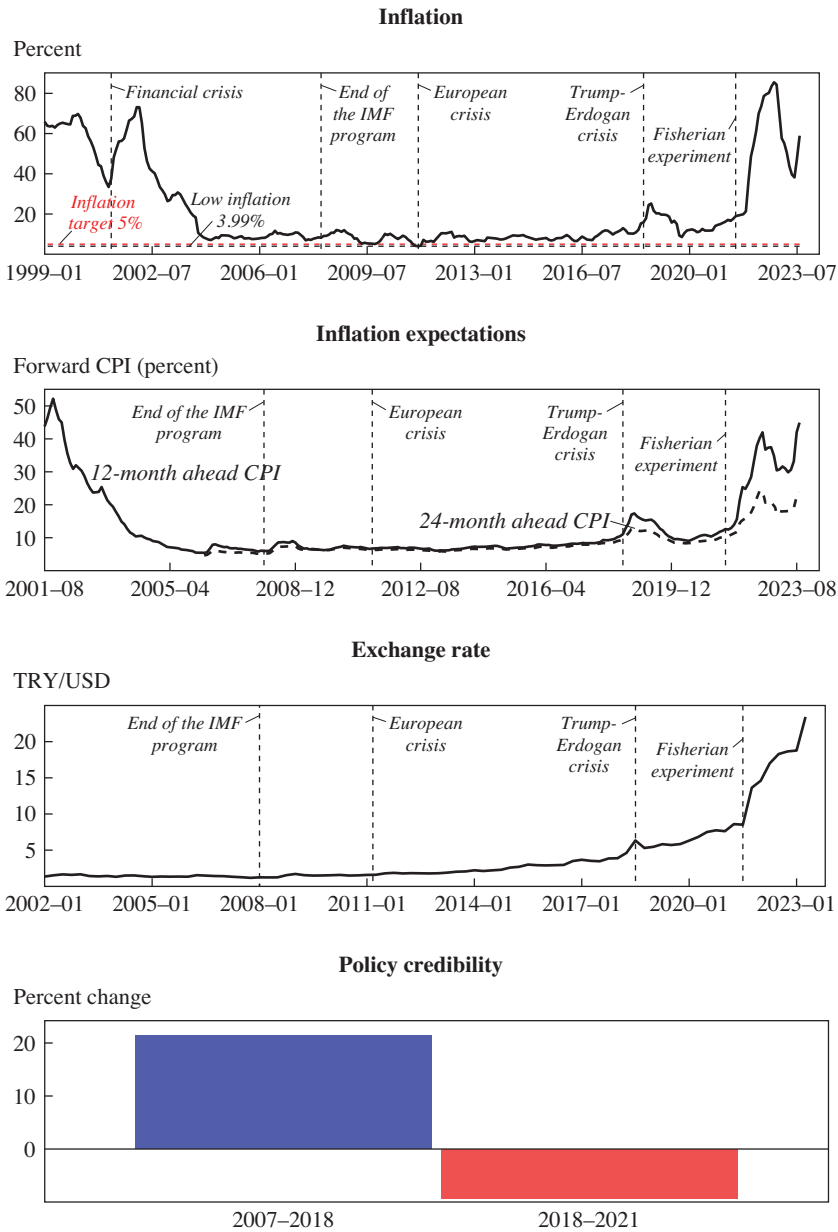
As shown in figure 5, the inflation and inflation expectations came down around 2004–2005 and stayed low (with inflation sometimes even below the target of 5 percent) until Turkey started an unorthodox monetary policy experiment, known as the Fisherian experiment, in late 2020.[14] This late period of 2018–2021 is when our credibility measure shows a deterioration

12. The UIP premium decline for Canada is explained by the fact that the interest rate differential term went down more than the expected appreciation since Canada did not change the policy rate at the time. Capital flows also showed different patterns: there were capital outflows from Mexico, whereas Canada received capital inflows (these results are available upon request).

13. Note that with a slight depreciation and an expected appreciation of the Canadian dollar, there is a slight increase in the UIP premium for Canada.

14. *Economist* (2020).

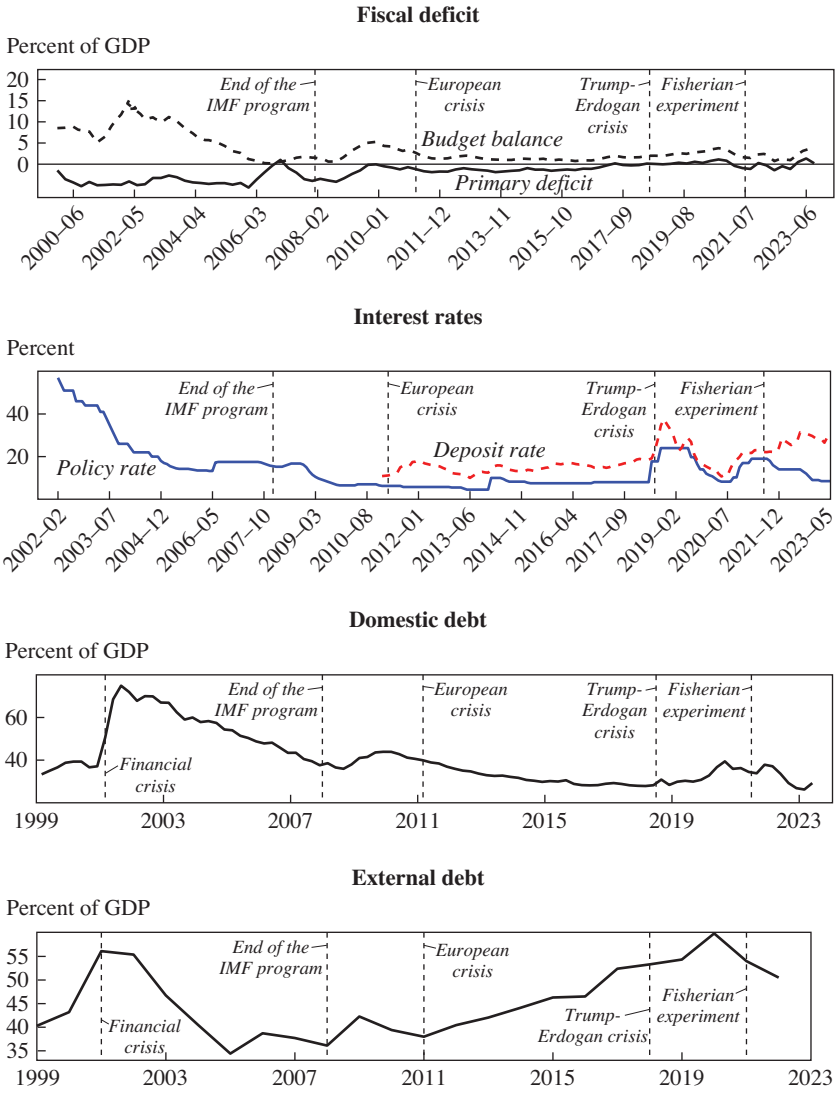
Figure 5. Case Study: Turkey I



Source: IMF International Financial Statistics (inflation and exchange rate); Central Bank of the Republic of Turkey Electronic Data Delivery System (EVDS) (inflation expectations); and IAPOC index (policy credibility).

Note: We plot exchange rates for the float regime starting in 2002.

Figure 6. Case Study: Turkey II



Source: Fiscal data come from Turkey's Ministry of Treasury and Finance; policy and deposit rate data are available from the Central Bank of the Republic of Turkey Electronic Data Delivery System (EVDS).

Note: The fiscal deficit is composed of primary deficit and budget balance—primary deficit data are the central government's last twelve-month ratio of primary balance to nominal GDP, and budget deficit data are calculated by adding the central government's last year ratio of interest expense share to primary deficit. Domestic debt as a percentage of GDP is the ratio of public sector net debt to GDP, covering total public gross debt stock, unemployment insurance fund net assets, public sector assets, and central bank net assets to last year's GDP. External debt as a percentage of GDP is the ratio of gross external debt stock to GDP, covering short- and long-term debt stocks of the public sector, the Central Bank of the Republic of Turkey, and the private sector.

of almost 10 percent, whereas the early period of 2007–2018 picks up an improvement of 20 percent (recall that the credibility index is between zero and one). In Turkey's case, the fluctuations in monetary policy credibility correlate increasingly well with inflation and inflation expectations, which act as lagging variables due to their nature as endogenous outcomes to changes in monetary policy credibility. Additionally, the nominal exchange rate depreciation, which began during the 2018 political crisis, further intensified in the later period, marked by a decline in policy credibility post-2020.[15]

Figure 6 shows the evolution of interest rates and domestic and external debt in Turkey. Again, the key insight here is not about the deteriorating fundamentals such as the current account deficit or external debt, as would typically be the case, but rather about how such deterioration priced in the risk spreads leads to different dynamics in market rates (short-term deposit rates) versus monetary policy rates, as shown to be the case in the latest episode.[16] Kalemli-Özcan (2019) calls this phenomenon "short-rate disconnect" and shows that emerging markets' domestic monetary policies have been ineffective in general since the 1990s as the policies' pass-through to domestic market rates is always less than one to one with capital flows having an effect on market rates as a function of risk sentiments. The Turkish case after 2020 is an example, with the monetary policy credibility deteriorating and priced in by foreign investors as a risk premium, which is picked up both by the UIP premia and as the difference between domestic market rates and policy rates. The issue is not only the less than one-to-one pass-through of policy rates into market rates, but also having these rates go in totally opposite directions. De Leo, Gopinath, and Kalemli-Özcan (2022) study the short-rate disconnect in detail by writing down a model that delivers the wedge between market rates and policy rates as long as the domestic financial intermediaries borrow overseas at a dollar premium. They show that emerging markets pursue countercyclical monetary policy; however, the market rates they face go up in bad times and down in good times due to the risk premia inherent in market rates for emerging markets, even though the monetary policy is countercyclical in those countries akin to advanced economies.

15. Tensions between Turkey and the United States soared as President Trump ordered new sanctions in 2018, following the political dispute over Turkey's continued detention of an American pastor who was jailed after a failed coup in Turkey. Tariffs on imported Turkish steel and aluminum were doubled to 50 percent and 20 percent, respectively (Tankersley, Swanson, and Phillips 2018).

16. We only plot external debt to save space as increasing external debt also implies widening current account deficits.

II. Data and Measurement

II.A. Monetary Policy Credibility

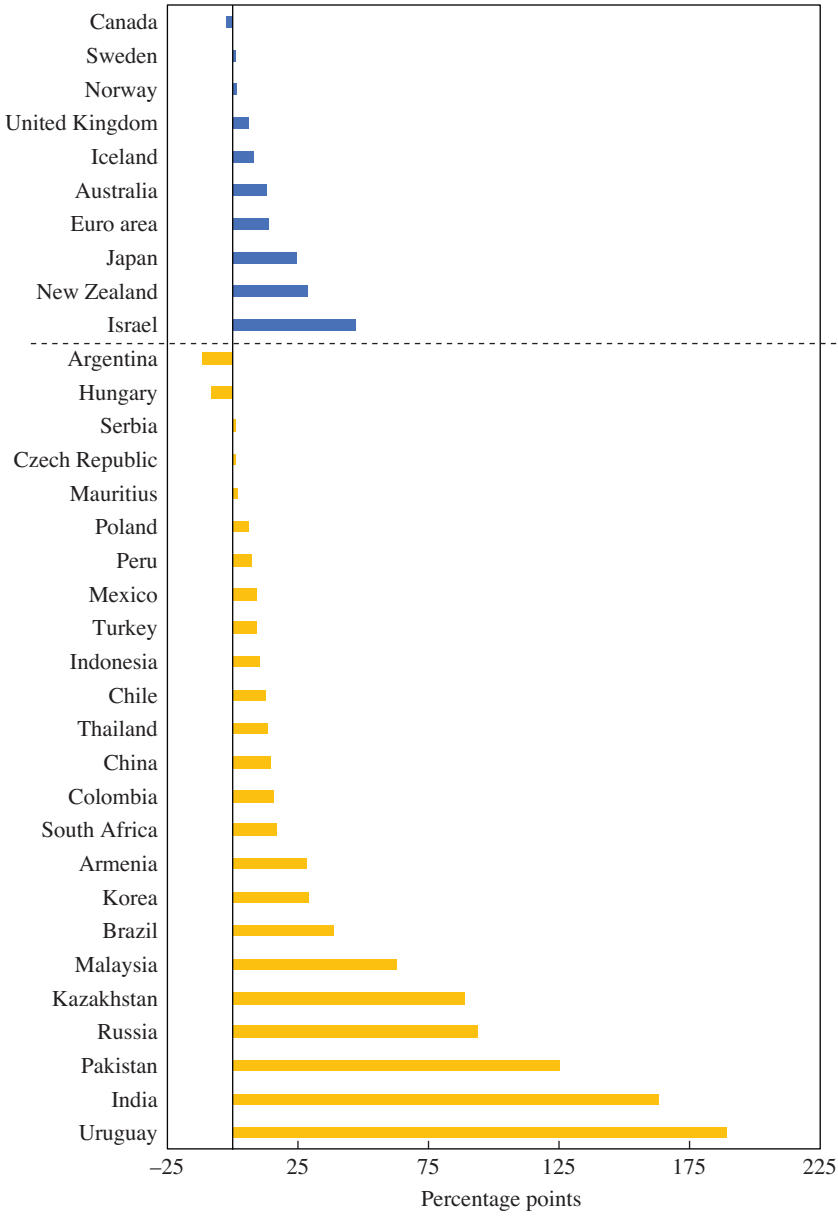
Our measure for monetary policy credibility is a new index developed by Unsal, Papageorgiou, and Garbers (2022) using a narrative approach similar to Romer and Romer (1989) for fifty countries between 2007 and 2021. This index characterizes monetary policy frameworks across three pillars: independence and accountability (IA), which provide the foundations of monetary policy; policy and operational strategy (PO), which guide the adjustments to policy stance given the objectives, as well as the adjustments to policy instruments to implement the policy stance; and communications (C), which conveys decisions about the policy stance and rationale to the public. To cover these pillars with sufficient clarity and comprehension, 225 criteria were used and assessed against the public information from countries' central banks. Figure 7 shows the detailed cross-country heterogeneity, where countries like Uruguay and India show the maximum improvement.

The improvement in monetary policy credibility becomes even more evident when comparing the distributions of the index for 2007 and 2021 in figure 8. The mass has shifted more to the right, keeping the extensive heterogeneity. Advanced economies have a narrower distribution. In particular, in 2007 for emerging markets, the lowest value is 0.194 and the highest is 0.759 (mean of 0.546). In the 2021 distributions, the highest value for emerging markets is 0.822, and the value for advanced economies is only 0.867; so the best monetary policy credibility in emerging markets is almost as good as the best among advanced economies.

The IAPOC index is negatively and significantly correlated with inflation and inflation expectations at different horizons (figure 9). The figure clearly shows that the downward slopes (higher policy credibility, lower inflation, and lower inflation expectations) are mostly driven by emerging markets and not by advanced economies. In fact, this is what makes our policy credibility index stand apart from a large number of existing studies that measure monetary policy credibility with realized inflation or inflation expectations, which are endogenous measures of policy credibility, since the inflation level and expectations might be driven by policy credibility as we show above.[17]

17. For example, Bems and others (2021) obtain policy credibility measure from inflation, relying on historical data.

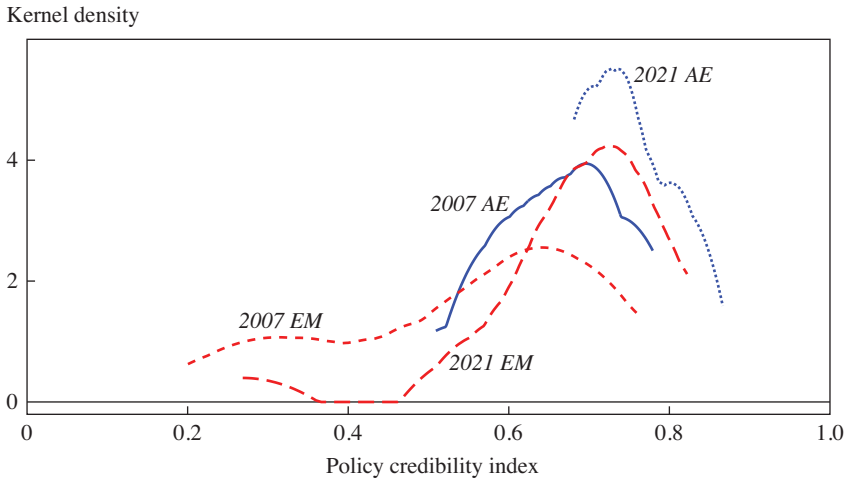
Figure 7. Change in Monetary Policy Credibility, 2007–2021



Source: IAPOC index from Unsal, Papageorgiou, and Garbers (2022).

Note: Percentage point change in monetary policy credibility of advanced economies and emerging markets between 2007 and 2021.

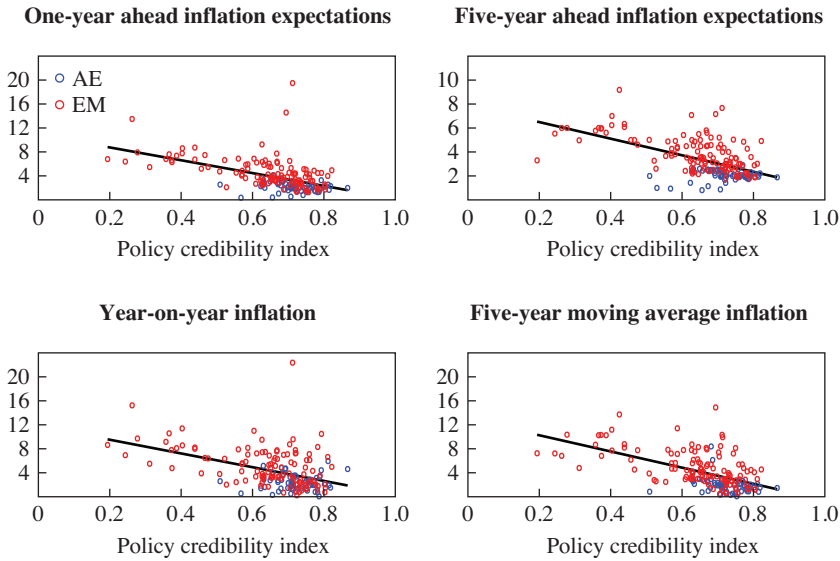
Figure 8. Policy Credibility Distributions



Source: IAPOC index from Unsal, Papageorgiou, and Garbers (2022).

Note: Distributions of policy credibility of advanced economies (AE) and emerging markets (EM) in 2007 and 2021.

Figure 9. Inflation and Expectations, 2007–2021



Source: IMF International Financial Statistics (inflation); IAPOC index from Unsal, Papageorgiou, and Garbers (2022); Consensus Forecasts and the *World Economic Outlook Projections* (April 2023 edition) (inflation expectations); and authors' calculations.

Note: Regression coefficients of one-year ahead inflation expectations, five-year ahead inflation expectations, year-on-year inflation, and five-year moving average inflation on policy credibility. Inflation is the headline CPI inflation and seasonally adjusted with ARIMA X-13.

II.B. Balance Sheet Weakness via FX Debt

To study the role of heterogeneity in terms of the balance sheet weakness of countries for the international transmission of US monetary policy, we rely on updated data from Fan and Kalemli-Özcan (2016) and Kalemli-Özcan, Liu, and Shim (2021) on the ratio of FX debt to total debt for the private sector in a given country, and we follow the methodology in Kalemli-Özcan (2019). These data come from the Bank for International Settlements (BIS) global liquidity indicators (GLI) database, which provides FX debt exposures for both bonds and loans for the nonfinancial private sector (nonfinancial corporations and households) and for governments separately. FX bonds are defined as debt securities issued in the US dollar, euro, or Japanese yen, and issued in international markets by the residents in the nonfinancial sector of a given economy. FX loans are defined as bank loans extended to the nonbank sector of a given economy by both domestic banks and international banks located outside the economy, and denominated in the US dollar, euro, or Japanese yen.

We work with the ratio of FX debt to total credit for the nonfinancial sector. Total credit data come from the BIS total credit database, which provides data on total loans and debt securities used for borrowing by the residents in the nonfinancial sector of a given economy, in both domestic and foreign currencies, and from both domestic and foreign lenders. By dividing the sum of loans and bonds in FX from the GLI data set for the nonfinancial sector by the sum of total loans and bonds for the nonfinancial sector from the total credit database, we obtain the country-level nonfinancial private sector FX debt share. The data are available for the following fifteen emerging economies: Argentina, Brazil, Chile, China, Colombia, India, Indonesia, Malaysia, Mexico, Peru, Philippines, Russia, South Africa, Thailand, and Turkey.

Of course, having FX debt alone does not necessarily indicate a weak balance sheet. To address this issue, we draw upon the extensive literature that documents how, in emerging markets, the financial sector (banks) is often required to hedge currency risk, while corporations, including exporters, tend not to match currency risk on their balance sheets (Di Giovanni and others 2022; Alfaro, Calani, and Varela 2023). Governments can act as the lender of last resort for dollars through their reserves, effectively hedging this risk at the national level, and hence we run robustness exercises controlling FX reserves, as reported in the online appendix figure A1.

The rationale for utilizing this data set, despite its limitations in terms of sample size, is its ability to focus exclusively on the private sector FX

exposure. This is crucial because, as we highlighted in the introduction, emerging market governments are increasingly borrowing in local currency. Even though we showed data from Bénétrix and others (2019) in the introduction, we do not use these data in our regressions as the FX dimension is a proxy in this data set. This is because it uses as input: the currency composition of the main international investment position (IIP) components from the International Monetary Fund (IMF); the IMF's Coordinated Portfolio Investment Survey (CPIS); the portfolio debt data reported to the European Central Bank; and banks' cross-border positions reported to the BIS, available through its locational banking statistics. Thus, corporate and government debt will be mixed, as those are mixed in the IIP and CPIS data sets, and hence the currency composition for the corporate sector cannot be precisely measured unlike our data from BIS.

II.C. Other Variables

Our panel data set includes other variables: GDP, Consumer Price Index (CPI), exchange rates, capital flows, and UIP deviations. We use seasonally adjusted real GDP from the *World Economic Outlook* and complement the missing series using data from central banks, national bureaus of statistics, and the International Financial Statistics (IFS). We use the CPI data from the IFS. For nominal exchange rates, we use the IFS as well. We also use total capital inflows, defined as the sum of bank, central bank, corporate, and government portfolio debt and other investment debt flows (loans) from BIS, originally constructed by Avdjiev and others (2022). These data are identical to the IMF balance of payments data at the annual level but with better quarterly coverage in emerging markets, which is why we prefer them over the standard IMF balance of payments data. The twelve-month UIP deviations are calculated as the difference between log interest rate differentials and the gap between log expected and spot exchange rate, all at the same horizon, as shown in section I. Log interest rate differentials are the short-term government bond rates vis-à-vis the United States, at twelve months. The log expected exchange rate is the twelve-month ahead expected exchange rate in a given month from the Consensus Economics, and the log exchange rate is the spot rate, both nominal and in terms of local currency per US dollar. From Bloomberg, we get the nominal interest rate data.

Our panel data set also includes other variables that we use as controls: trade balance to GDP, dollar shock, oil price index, and FX reserves to GDP. Data on trade balance to GDP are from the IFS. As for dollar shock,

Table 1. Country Sample

	Emerging markets	
Advanced economies		*Countries for which we have a direct measure of FX debt exposure of the private sector*
Australia	Albania*	Argentina
Canada	Armenia	Brazil
Denmark*	Azerbaijan*	Chile
Euro Area	Belarus*	China
Finland*	Bulgaria*	Colombia
Germany*	Costa Rica*	India
Iceland	Croatia*	Indonesia
Ireland*	Czech Republic	Malaysia
Israel	Ecuador*	Mexico
Italy*	Egypt Arab*	Peru
Japan	Guatemala*	Philippines
New Zealand	Hungary	Russia
Norway	Kazakhstan	South Africa
Spain*	Korea*	Thailand
Sweden	Latvia*	Turkey
Switzerland*	Malta*	
United Kingdom	Mauritius	
	Morocco*	
	Pakistan	
	Paraguay	
	Poland	
	Romania*	
	Serbia	
	Singapore*	
	Slovak Republic*	
	Tunisia*	
	Uruguay	

Source: Authors' compilation.
* Indicates no IAPOC index measure for this country.

we use the Nominal Major Currencies US Dollar Index from FRED, and we normalize it to 10 percent following Obstfeld and Zhou (2022). Oil prices and FX reserves to GDP data are from the IFS. In our analysis, we drop hard pegs and dual markets exchange rate countries (Ilzetzki, Reinhart, and Rogoff [2022] classifications 1 and 6). Thus, we always work with an unbalanced panel composed of managed and pure floats at the time of their inclusion.

Table 1 lists our country sample. We have a total of fifty-nine countries in the big sample. These are all advanced economies and emerging markets that do not have hard pegs and dual markets exchange rates. Similarly, of the fifty countries that are in the IAPOC index sample, we work with thirty-four; we drop the low-income countries, those with hard pegs, dual

markets exchange rate countries, and the United States. In the FX debt exercise, we have only fifteen emerging economies, all floating or managed floating countries. The online appendix provides more details including descriptive statistics.

III. Empirical Analysis

III.A. Fed Hikes and Risk Premia in Financial Markets

We want to capture the exogenous component of US monetary policy that constitutes a surprise for the financial markets, which in turn has an impact on their risk sentiment, after a Federal Reserve announcement. Not every Fed hike needs to involve a change in the risk sentiments of investors, but if there are enough Fed hikes that do change the risk sentiments, then our identification of the risk channel of US monetary policy's international transmission is valid. We are also relying on the fact that a large body of literature shows a high correlation between the Fed hikes and common measures of risk sentiments (e.g., the VIX and the excess bond premium). We also use such measures for robustness in addition to our exogenous US monetary policy measures.[18]

The US monetary policy is endogenous to the US business cycle and financial markets since markets price in the expected actions of the Federal Reserve before the actual change in the policy rate. The common approach to dealing with the endogeneity of monetary policy in the literature is to measure the monetary policy surprises. These surprises are obtained from high-frequency changes in interest rates around central bank policy announcements. The key identifying assumption is that the monetary policy is predetermined over the event window and hence not affected by the financial market reaction. Using such surprises, the macro finance literature estimates the causal effect of US monetary policy both on financial markets (Kuttner 2001; Gürkaynak, Sack, and Swanson 2004) and on macro variables (Stock and Watson 2018; Gertler and Karadi 2015).

Recently, this literature has been debating some puzzling effects. Forecasts respond in the wrong direction when a high-frequency monetary policy surprise indicates, say, a tightening of monetary policy. Not only do output, employment, and inflation respond positively to tightening (Nakamura and Steinsson 2018), but similar positive responses are observed

18. Results with the VIX, excess bond premium, and a new measure of risk-on-risk-off (RORO) sentiment from Chari, Dilts Stedman, and Lundblad (2020) are available upon request.

in the stock market as well (Miranda-Agrippino and Ricco 2023; Cieslak and Schrimpf 2019; Jarociński and Karadi 2020). The common explanation for these puzzling results is the "Federal Reserve information effect," that is, the Federal Reserve announcements convey private information about the economy and therefore directly affect the beliefs about economic fundamentals. If, for example, a tightening surprise is interpreted as a signal that the Federal Reserve thinks the economy is stronger, then the survey forecasters will revise their outlook upward and the stock market will boom. As a result, monetary policy surprises are not exogenous but contaminated with information that will prevent them from identifying the causal effects of monetary policy.

There is also the additional problem of relevance. This problem is about the fact that the surprises are small. In fact, Obstfeld and Zhou (2022) argue that the US dollar exchange rate is a better measure than the monetary policy shocks for tracing the risk-based international transmission from the United States to the rest of the world, since the dollar exchange rate picks up much more variation in risk sentiment variables such as the VIX and the excess bond premium. Consistently, others argue that the most important driver of the global financial cycle is not the US monetary policy per se, but rather the precise measures of risk sentiments such as the excess bond premium (Rogers, Sun, and Wu 2023) and volatility in macroeconomic news (Boehm and Kroner 2023). Unfortunately, all of these—the dollar exchange rate, VIX, excess bond premium, and macroeconomic news—are endogenous to the US monetary policy changes since they are all endogenous to financial markets' risk sentiment changes that depend largely on US monetary policy.

For example, when the Federal Reserve hikes the rates, the global financial conditions get tighter, which results in a higher excess bond premium, flight to safety, and an appreciation of the US dollar together with more macroeconomic news on higher earning volatility and uncertain outlook. For our purposes, we want the US monetary policy surprises that are exogenous to the US economy and financial markets but still relevant for financial markets, relevant enough that the surprises will change financial markets' risk sentiments. We do not want our policy surprises to be contaminated by the Federal Reserve or the financial markets' reaction to public news that is available before the Federal Reserve announcement. Rather, we want to measure the new information that financial markets learn from the Federal Reserve's announcement and changes their risk sentiments and international portfolios differentially across emerging markets versus advanced economies.

Table 2. Weak Instrument Test

	Cragg-Donald	*Wald* F *statistic*	*Kleibergen-Paap* rk	*Wald* F *statistic*
Depvar	*Emerging markets*	*Advanced economies*	*Emerging markets*	*Advanced economies*
GDP	370.261	248.115	370.297	248.320
Capital inflows to GDP	175.319	74.783	175.251	74.716
Exchange rate	440.293	257.478	440.532	257.772
Twelve-month UIP deviation	144.371	111.145	144.376	111.096

Source: Authors' calculations.

Note: Shown are the weak instrument test results for the baseline regression (specification one below) and for $h = 1$. They are all above the Stock-Yogo weak ID test critical values of 10 percent maximal IV size, which in this case is equal to 16.38.

Bauer and Swanson (2023) solve these types of endogeneity issues. They show that the key endogeneity problem lies in the omitted variable of economic news, where all—survey forecasters, markets, and the Federal Reserve policy—respond to macroeconomic news. Bauer and Swanson (2023) show that there is no information effect in the Federal Reserve's announcements, but rather that the predictability of the monetary policy surprises is due to learning about the Federal Reserve's policy during the announcements. Hence, the publicly observable macro data and the omitted news can help solve the endogeneity issue together with the relevance issue. Bauer and Swanson (2023) compute the orthogonalized monetary surprises as residuals from regressing monetary surprises on six macro and financial variables. As a result, we use monetary policy surprises from both Gertler and Karadi (2015) and Bauer and Swanson (2023) in our analysis. We use Gertler and Karadi (2015) in a two-step IV approach using the surprises, calculated as the movements in the prices of short maturity (three-month) federal funds futures contract in a thirty-minute window surrounding the Federal Open Market Committee announcement, as instruments for the policy rate (the twelve-month T-bill rate). We use Bauer and Swanson (2023) surprises in reduced form. Following Bauer, Bernanke, and Milstein (2023), we rescale the Bauer and Swanson (2023) surprises to gauge the effects of a 10 basis point surprise (the standard deviation of the original surprises is about 9 basis points).

The monetary policy shocks from Gertler and Karadi (2015) comfortably pass the weak instrument tests, and hence they are relevant in capturing the exogenous changes in US monetary policy, as we show in table 2 (regressions of the US policy rate on policy surprises).

III.B. Historical Evidence: The Impact of Fed Hikes on Emerging Markets versus Advanced Economies, 1990:Q1–2019:Q4

To uncover the asymmetric effects of Fed hikes, we rely on local projections, as proposed by Jordà (2005). The local projection method provides a flexible framework and is easy to implement. Moreover, it is well documented that local projections have several advantages over the vector autoregression (VAR) models. Above all, local projections are more robust to possible misspecifications, at least under a finite lag structure (Kilian and Lütkepohl 2017; Plagborg-Møller and Wolf 2021). They allow us to parsimoniously model the asymmetric effects of US monetary policy on emerging markets versus advanced economies, on countries with high versus low policy credibility, and also on countries with high versus low debt denominated in US dollars. The local projections estimation also saves degrees of freedom relative to a multivariate approach: even though we lose observations from adjusting for leads and lags, our set of control variables on the right-hand side is relatively sparse as we do not need to describe the dynamics of the endogenous variables conditional on the shock.

Local projections regress the dependent variable at different horizons $t + h$ for $h = 1, 2, \ldots H$, conditional on an information set that consists of a set of control variables. In the linear case, the regression equation reads:

$$y_{t+h} = \alpha_h + \beta_h Shock_t + \gamma X_t + \varepsilon_{t+h},$$

where y_{t+h} is the variable of interest at horizon h and X_t is a vector of control variables, contemporaneous and lagged as long as they are supposed to have an effect on the endogenous variable y_{t+h}, independent from the identified structural shock, $Shock_t$.

These control variables in X_t deserve discussion. The international transmission literature uses the specification below in general (Rey 2013; Degasperi, Hong, and Ricco 2023; Miranda-Agrippino and Rey 2020; Kalemli-Özcan 2019):

$$(1) \qquad y_{c,t+h} = \alpha_c + \beta_h \hat{\imath}_t^{US} + \sum_{i=1}^{i=4} \omega_i X_{t-i} + \sum_{i=1}^{i=4} \eta_i x_{c,t-i} + \varepsilon_{c,t+h},$$

where $y_{c,t+h}$ is a vector of macro and financial variables of country c at horizon h and α_c are country fixed effects that absorb institutional differences across countries, including slow-moving fundamentals.

There are two sets of controls, all of which enter lagged: X_{t-i} are lags of the global controls for the shock (lags of monetary policy rate, $\hat{\imath}_t^{US}$, and lags of monetary policy surprises that instrument the policy rate); and $x_{c,t-i}$ are lags of dependent variable and lags of country-specific controls that have an independent effect but are correlated with the past and anticipated US policy changes. These are inflation rate differentials and GDP growth differentials for the given country with the United States. These controls are essential since the inflation rate differentials are key for the financial channel of policy transmission, and GDP growth differentials are key for the trade channel. Investors switching demand for assets or consumers switching demand for goods between countries as a result of the past or anticipated changes in US policy and other global shocks are captured directly by these variables.

What then remains to be captured by the identified US monetary policy shock is the transmission via the financial channel driven by endogenous changes in the risk premium affecting the current and future interest rate differentials. Policy transmission via the trade channel will be captured by the endogenous appreciation of the dollar affecting the current and future GDP growth differentials. We investigate the impact of identified US shocks on both risk premia and exchange rates. When $y_{c,t+h}$ is GDP and shows improvement, the trade channel should be dominant; whereas, if GDP deteriorates, then the financial channel is the dominant channel of international transmission. Notice that two of the other endogenous outcomes we focus on—capital flows and exchange rates—cannot separate the channels of transmission since both channels will imply capital flows out on net (or net exports increase) and exchange rate depreciates vis-à-vis the dollar. But the falling GDP and rising risk premia (UIP) can identify the financial channel dominating over the trade channel.

Last but not least, $\hat{\imath}_t^{US}$ denotes the instrumented twelve-month US Treasury rate, where the first stage regresses the Treasury rate on monetary policy surprises from the three-month federal funds futures contract prices, following Gertler and Karadi (2015) as we explained in the previous section. As we also showed before, the instrument passes the relevance test, meaning the Gertler-Karadi shocks we use are not weak instruments for the US monetary policy changes.

Although we believe that the parsimonious specification given in equation (1) is all that is needed to identify the asymmetric effects of US policy on emerging markets versus advanced economies, to ease the worries about robustness, we also run equation (2) to control for additional global variables contemporaneously. This exercise will show that we do not

need to control for additional variables as none of our results based on equation (1) will change qualitatively, and conditional on the equation (1) variables, additional variables from equation (2) will not have much explanatory power.

For this exercise, we follow Obstfeld and Zhou (2022) and run the following specification with additional global controls, allowing both contemporaneous and lagged relation between these variables and the identified US monetary policy shock:

$$(2) \qquad y_{c,t+h} = \alpha_c + \beta_h \hat{\imath}_t^{US} + \gamma X_t + \sum_{i=1}^{i=4} \omega_i X_{t-i} + \sum_{i=1}^{i=4} \eta_i x_{c,t-i} + \varepsilon_{c,t+h}.$$

The variable X_t is a vector of global controls including the US dollar shock from Obstfeld and Zhou (2022), defined as the appreciation of the US dollar vis-à-vis euro area, Canada, Japan, United Kingdom, Switzerland, Australia, and Sweden, the oil price index, and the median country trade balance. When we run regressions for emerging markets and advanced economies separately, we use the median trade balances specific to those aggregate groups. The variable X_{t-i} includes the lags of all these global controls.

III.C. Benchmark Results

Figure 10 displays the differential impact of the US monetary tightening on advanced economies and emerging markets, based on equation (1) where we run this in the two samples of countries. The US monetary policy shock results in a significant and persistent decline in output in emerging markets but not in advanced economies: a 1 percentage point increase in the US policy rate leads to a 2 percent decline in output by the third quarter and a 3 percent decline by the ninth quarter in emerging markets. The stark difference between the output results implies that the financial channel dominates the trade channel in emerging markets.

The dominance of the financial channel of US policy transmission for emerging markets can also be seen from the large nominal exchange rate depreciation observed in quarters two to four (whereas advanced economies' exchange rates do not respond significantly) combined with the large increase in UIP: 3.5 percentage points for a 1 percentage point shock by the third quarter. Given the mean UIP deviation for emerging markets, this implies a large change: moving from a country that is in the 25th percentile to a country in the 75th percentile of the UIP wedge distribution, which would be moving from Chile to Argentina. Recall that a higher UIP premium means higher expected excess returns to local currency vis-à-vis the dollar.

Figure 10. International Transmission of the Fed Hikes: Emerging versus Advanced Economies (Gertler-Karadi Surprises)

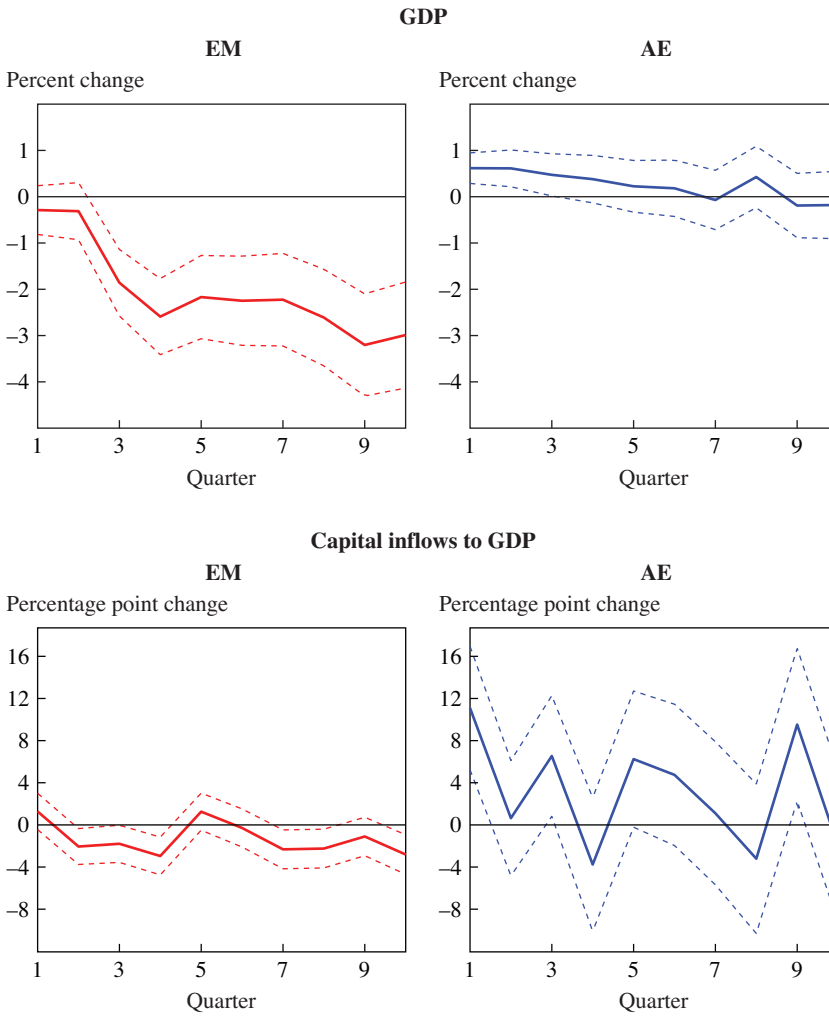
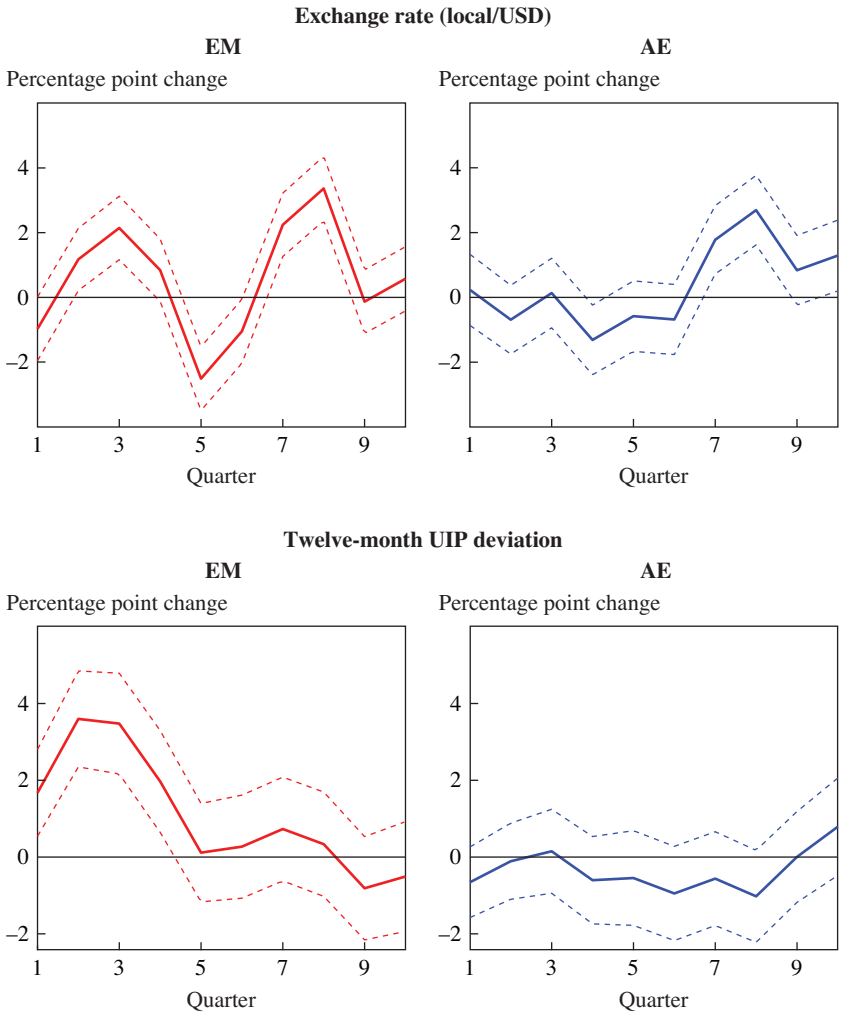


Figure 10. International Transmission of the Fed Hikes: Emerging versus Advanced Economies (Gertler-Karadi Surprises) (*Continued*)



Source: Authors' calculations.

Note: Impulse responses of the twelve-month US Treasury rate, instrumented by monthly weighted raw surprises in the three-month federal funds futures from Gertler and Karadi (2015), are obtained from panel local projections. Confidence intervals at 90 percent (calculated using Newey-West standard errors) are indicated by the dashed lines. Controls include four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, and the instrument. See also figure A1 in the online appendix, where we add FX reserves to GDP as a control and where the advanced economies' exchange rates also show some depreciation. Dependent variables include real GDP in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), UIP deviations, which are defined as the twelve-month interest rate (government bond) differentials vis-à-vis the United States minus the expected changes in the exchange rate, and the ratio of total capital inflows to GDP. See also figure A2 in the online appendix, where we also run this specification for our smallest country sample (FX debt EM sample).

It can happen if investors expect the emerging market's currency to appreciate in the future since there is a depreciation on impact with the Fed hike, or the emerging market's interest rate differentials with the United States increase as a result of higher risk premium, or both.[19] Consistent with higher UIP premia, capital inflows go down (meaning international investors leave) by 2 percentage points around the third quarter before reverting back. All these variables are insignificant for advanced economies.

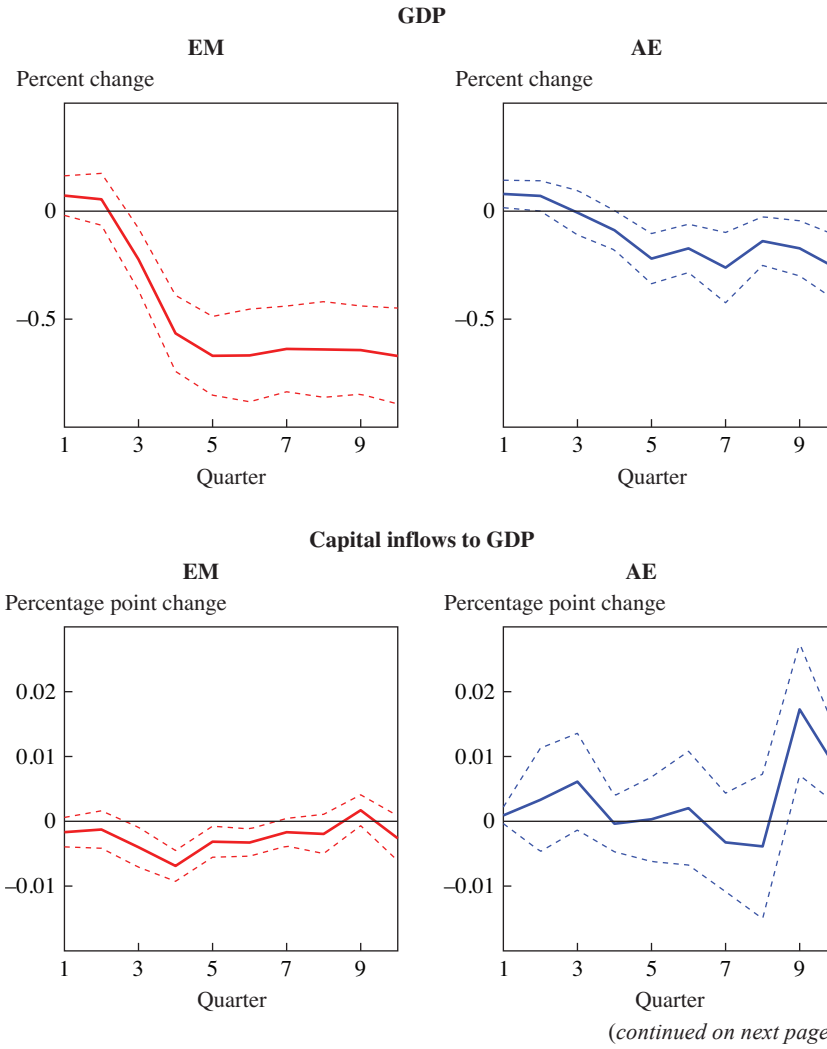
We next run equation (1) in reduced form, using the monetary policy surprises in Bauer and Swanson (2023). Figure 11 shows results that are similar for emerging markets with more significant capital outflows. In particular, a 10 basis point shock results in a 0.2 percent decline in output by the third quarter and 0.6 by the ninth quarter in emerging markets. Similarly, the dominance of the financial channel is shown by an increase in UIP of 0.8 percentage points by the third quarter for emerging markets, while there is no effect at all for advanced economies. What is interesting is that now we also have a decline in output for advanced economies combined with currency depreciation. Hence, even for advanced economies, the financial channel dominates the trade channel, but the impact is much milder on output since there is no response of UIP wedge and capital outflows to the US shocks in advanced economies.

In figure 12, we show the results of equation (2), which includes global controls that might be correlated with the US policy shocks. Results are consistent with our previous findings. In figure A3 in the online appendix, we rerun this exercise, dropping commodity exporters, and find that the results hold with the exception that now we also have some delayed depreciation in the advanced economies' exchange rates.

In figure 13, we show the results of running equation (2) in reduced form using the monetary policy shocks from Bauer and Swanson (2023). We do not find large differences relative to our findings in figure 11, which highlights the strength of the results. The only change is that now the previous, mild decline on advanced economies' GDP goes away, and in fact, there is a weak small increase in GDP together with currency depreciation, which would support the trade channel via expenditure switching. The problem is that by the third quarter, when currency depreciates, the output effect becomes insignificant.

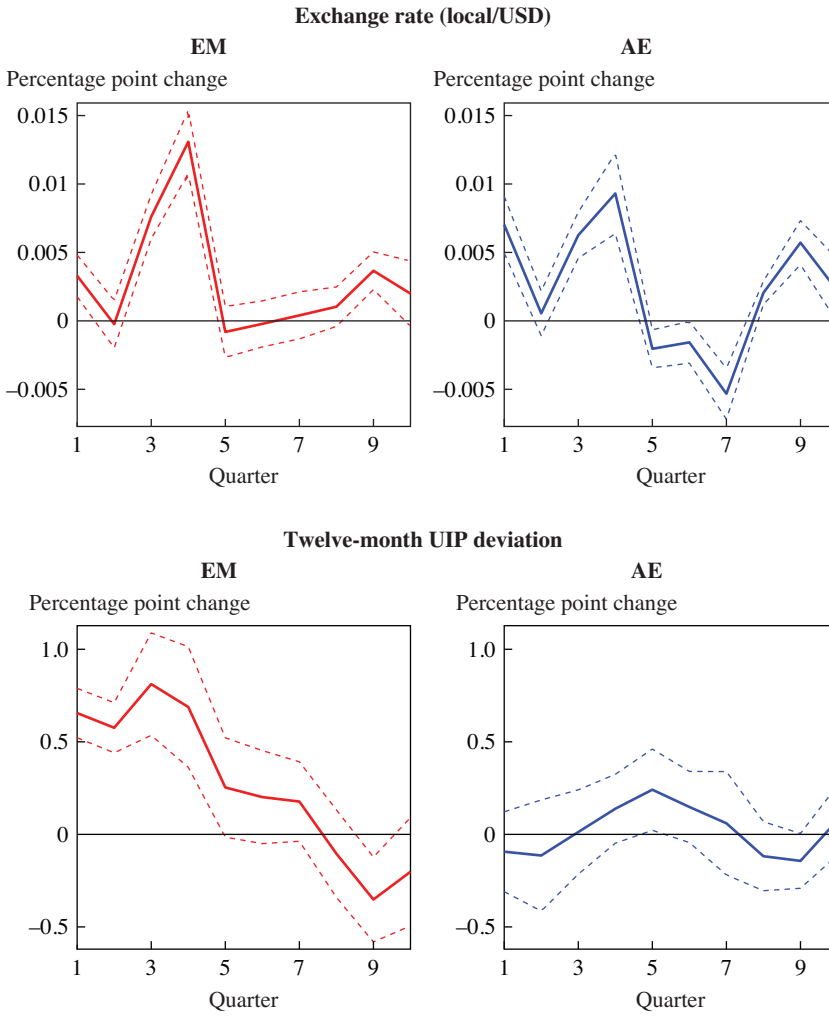
19. This result is not due to higher policy rates in emerging markets, as shown by De Leo, Gopinath, and Kalemli-Özcan (2022).

Figure 11. International Transmission of the Fed Hikes: Emerging versus Advanced Economies (Bauer-Swanson Surprises)



(*continued on next page*)

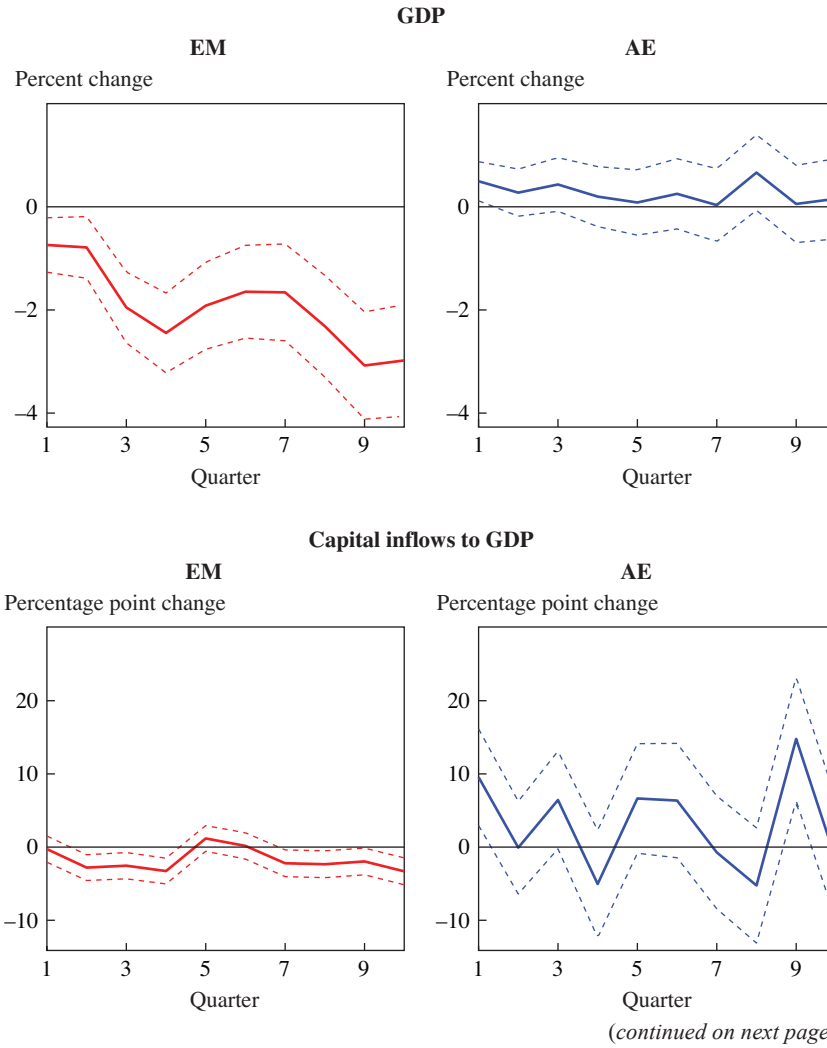
Figure 11. International Transmission of the Fed Hikes: Emerging versus Advanced Economies (Bauer-Swanson Surprises) (*Continued*)



Source: Authors' calculations.

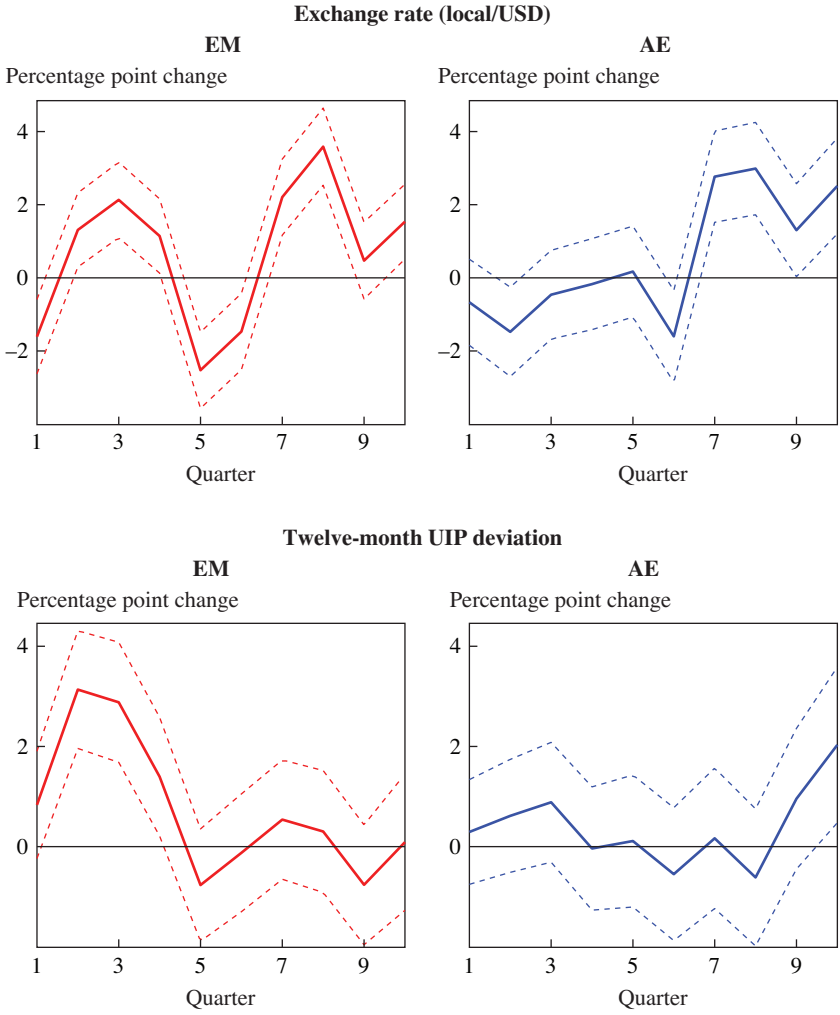
Note: Impulse responses of the US monetary policy surprises in Bauer and Swanson (2023), scaled to a 10 basis point surprise, are obtained from panel local projections. Confidence intervals of 90 percent (calculated using Newey-West standard errors) are indicated by the dashed lines. Controls include four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, and the shock. Dependent variables include real GDP in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), UIP deviations, which are defined as the twelve-month interest rate (government bond) differentials vis-à-vis the United States minus the expected changes in the exchange rate, and the ratio of total capital inflows to GDP.

Figure 12. International Transmission of the Fed Hikes: Emerging versus Advanced Economies with Global Controls (Gertler-Karadi Surprises)



(*continued on next page*)

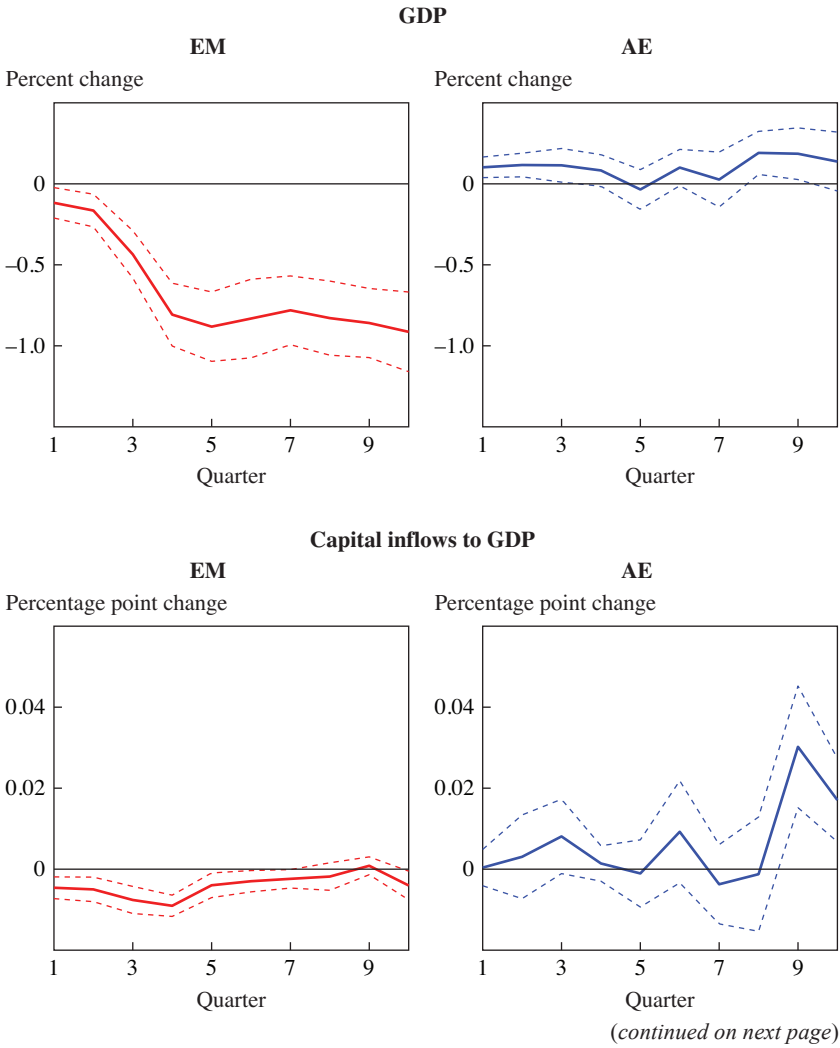
Figure 12. International Transmission of the Fed Hikes: Emerging versus Advanced Economies with Global Controls (Gertler-Karadi Surprises) (*Continued*)



Source: Authors' calculations.

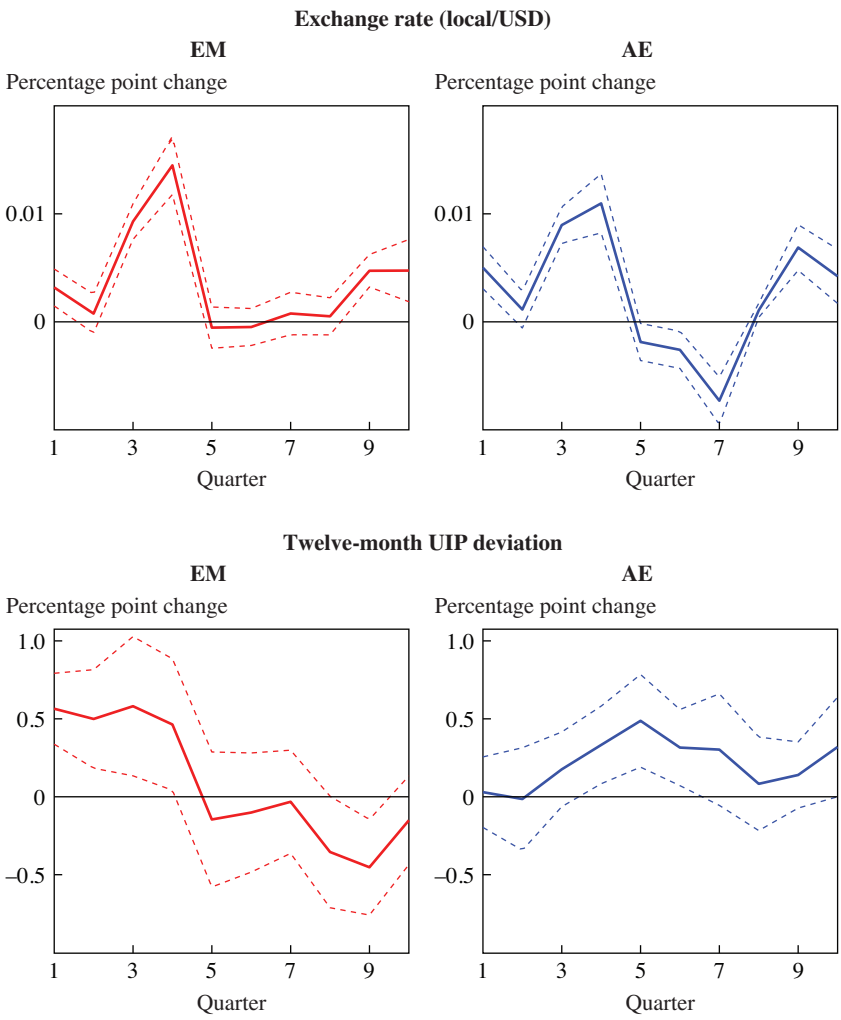
Note: Impulse responses of the twelve-month US Treasury rate, instrumented by monthly weighted raw surprises in the three-month federal funds futures from Gertler and Karadi (2015), are obtained from panel local projections. Confidence intervals at 90 percent (calculated using Newey-West standard errors) are indicated by the dashed lines. Controls include four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, the instrument, dollar shock, average oil price index, and median trade balance. Global controls (the last three) also enter contemporaneously. Dependent variables include real GDP in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), UIP deviations, which are defined as the twelve-month interest rate (government bond) differentials vis-à-vis the United States minus the expected changes in the exchange rate; and the ratio of total capital inflows to GDP.

Figure 13. International Transmission of the Fed Hikes: Emerging versus Advanced Economies with Global Controls (Bauer-Swanson Surprises)



(*continued on next page*)

Figure 13. International Transmission of the Fed Hikes: Emerging versus Advanced Economies with Global Controls (Bauer-Swanson Surprises) (*Continued*)



Source: Authors' calculations.

Note: Impulse responses of the US monetary policy surprises in Bauer and Swanson (2023), scaled to a 10 basis point surprise, are obtained from panel local projections. Confidence intervals of 90 percent (calculated using Newey-West standard errors) are indicated by the dashed lines. Controls include four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, monetary policy shocks, dollar shock, average oil price index, and median trade balance. Global controls (the last three) also enter contemporaneously. Dependent variables include real GDP in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), UIP deviations, which are defined as twelve-month interest rate (government bond) differentials vis-à-vis the United States minus the expected changes in the exchange rate, and the ratio of total capital inflows to GDP.

III.D. The Role of Policy Credibility

Why are emerging markets affected worse from Fed hikes (at least historically, during the period we study: 1990:Q1–2019:Q4)? To shed light on this question, we extend our local projections framework to analyze the differential impact of the US monetary policy shocks depending on the monetary policy credibility of countries, where we rely on the IAPOC index by Unsal, Papageorgiou, and Garbers (2022). In particular, we augment equation (2) in the following way:

$$y_{c,t+h} = \alpha_c + \beta_{1,h}\hat{\imath}_t^{US} + \beta_{2,h}\hat{\imath}_t^{US} * IAPOC_{c,2007} + \gamma X_t + \sum_{i=1}^{i=4} \omega_i X_{t-i} + \sum_{i=1}^{i=4} \eta_i x_{c,t-i} + \varepsilon_{c,t+h}, \tag{3}$$

where $IAPOC_{c,2007}$ is time in-varying and takes the 2007 initial value for each country.

To calculate the effect of the US monetary policy shock on countries with high versus low policy credibility, we calculate the marginal effect of a US monetary policy shock as:

$$\frac{\partial_y}{\partial_{\hat{\imath}}} = \beta_{1,h} + \beta_{2,h} * IAPOC_{2007}, \tag{4}$$

and we evaluate equation (4) at the 25th percentile of the 2007 IAPOC index distribution for the low-credibility country and at the 75th percentile for the high-credibility country.

Figure 14 shows the impulse response functions, which are striking. As shown, countries with low monetary policy credibility experience sharper contractions in output and higher UIP deviations, even though the extent of nominal exchange rate depreciations is similar among low and high credibility countries. We also plot inflation response where, interestingly, the low credibility countries have declining inflation, reflecting the severe contraction of the economy. In fact, given the high exchange rate pass-through in countries with low credibility, it can be that the central banks increase interest rates, which would further slow down growth and increase the UIP wedge. Instead, central banks with high credibility can afford to support the economy by lowering interest rates after the shock.

III.E. The Role of Balance Sheet FX Vulnerabilities

Another reason why emerging markets were affected worse from Fed hikes historically can be their sizable external debt that is financed with

Figure 14. International Transmission of the Fed Hikes: The Role of Policy Credibility with Global Controls (Gertler-Karadi Surprises)

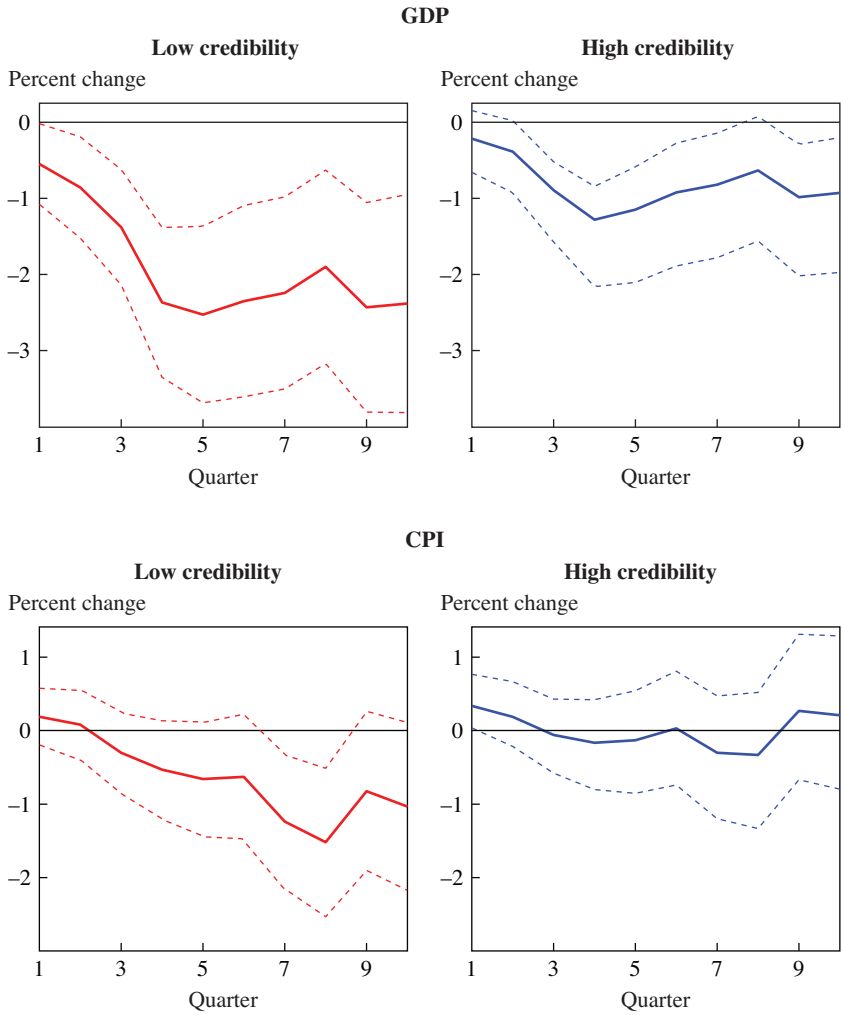
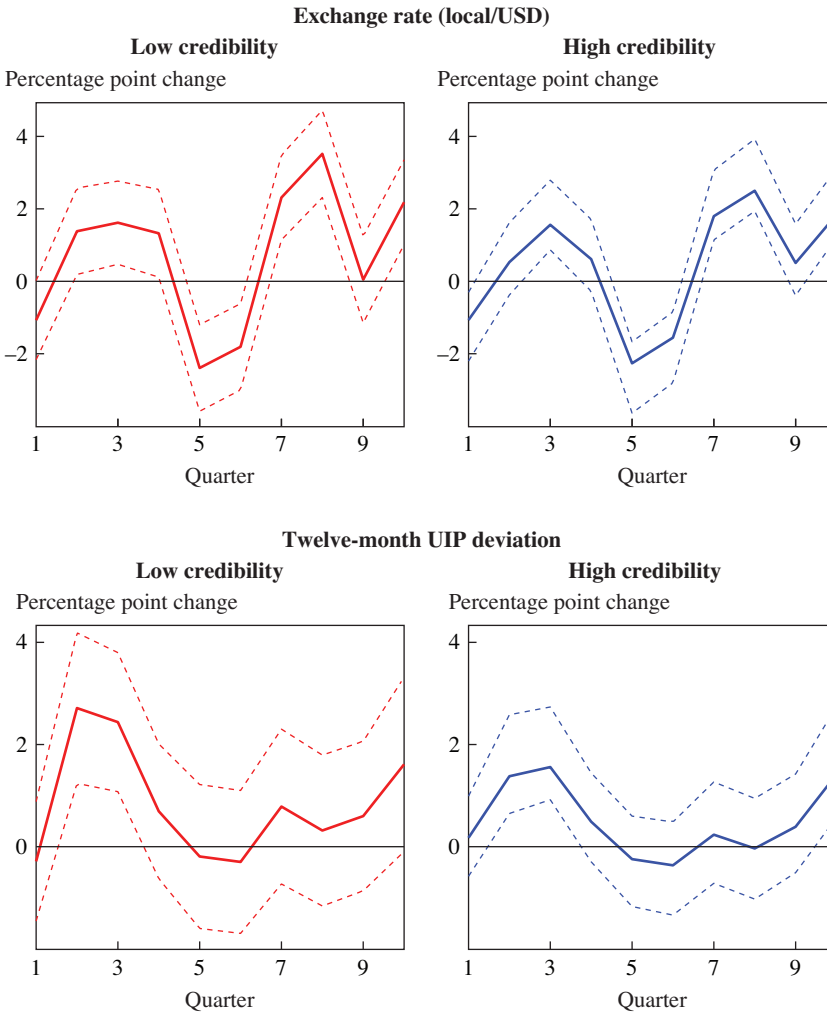


Figure 14. International Transmission of the Fed Hikes: The Role of Policy Credibility with Global Controls (Gertler-Karadi Surprises) (*Continued*)



Source: Authors' calculations.

Note: Impulse responses of the twelve-month US Treasury rate, instrumented by monthly weighted raw surprises in the three-month federal funds futures from Gertler and Karadi (2015), are obtained from panel local projections. Confidence intervals at 90 percent (calculated using Newey-West standard errors) are indicated by the dashed lines. Controls include four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, the instrument, dollar shock, average oil price index, and median trade balance. Global controls (the last three) also enter contemporaneously. Dependent variables include real GDP in logs, CPI in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), and UIP deviations, which are defined as the twelve-month interest rate (government bond) differentials vis-à-vis the United States minus the expected changes in the exchange rate. See text for the definitions of high and low credibility countries.

persistent current account deficits and largely denominated in US dollars. Such debt creates balance sheet vulnerabilities hindering investment and growth, especially when the cost of servicing this debt goes up with Fed hikes where assets on balance sheets are largely in local currency, as shown by Kalemli-Özcan (2019).

We extend our local projections framework to allow the impact of the US monetary policy shocks to differ based on FX (US dollar) debt of the private nonfinancial sector. We augment our equation (2) in the following way:

$$y_{c,t+h} = \alpha_c + \beta_{1,h}\hat{\imath}_t^{US} + \beta_{2,h}\hat{\imath}_t^{US} * FXdebt_{c,2000} + \gamma X_t + \sum_{i=1}^{i=4} \omega_i X_{t-i} + \sum_{i=1}^{i=4} \eta_i x_{c,t-i} + \varepsilon_{c,t+h}, \tag{5}$$

where $FXdebt_{c,2000}$ is a time-invariant variable equal to the initial 2000 value of FX debt.

To calculate the effect of the US monetary policy shock on high versus low FX debt countries, we calculate the marginal effect of a US monetary policy shock as:

$$\frac{\partial_y}{\partial_{\hat{\imath}}} = \beta_{1,h} + \beta_{2,h} * FXdebt_{2000}. \tag{6}$$

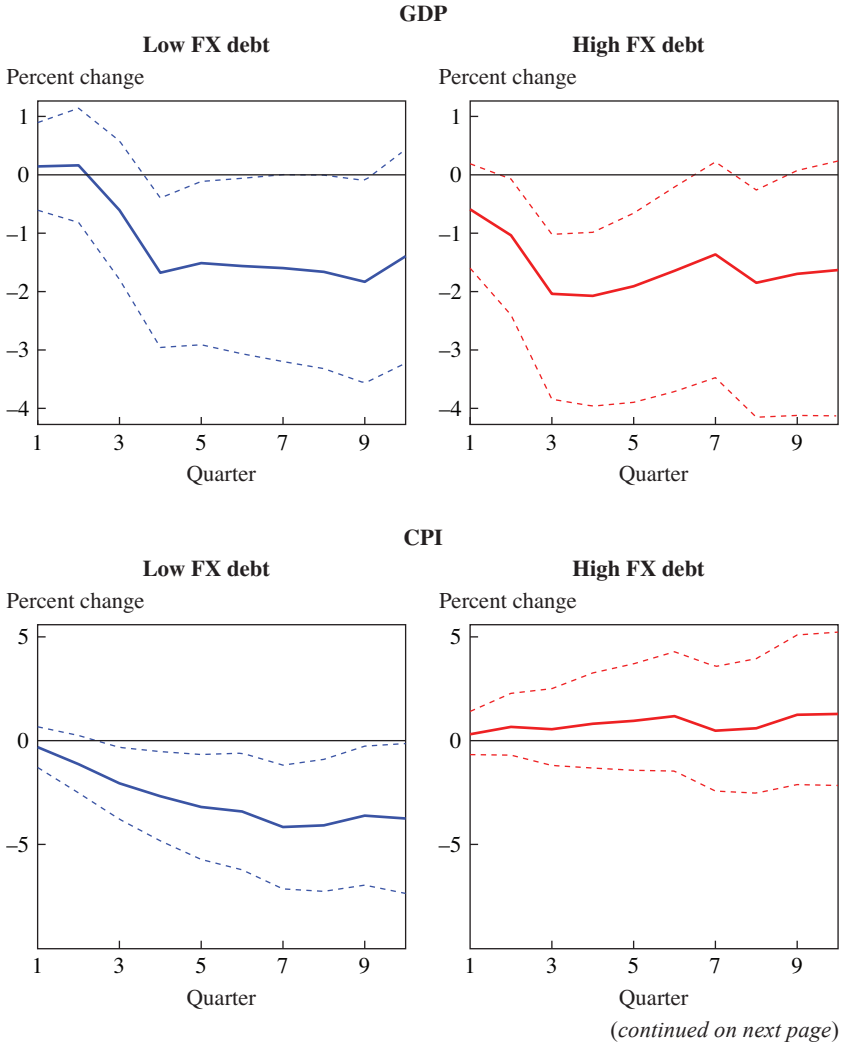
For the low FX debt country, we evaluate equation (6) using the minimum value of the 2000 FX debt distribution; and for the high FX debt country, we evaluate the same equation using the maximum value of that initial distribution.

We summarize the impulse response functions in figure 15. Countries with high FX debt go through sharper contractions in output on impact together with longer depreciations, higher inflation, and capital outflows, though given the small sample size, the statistical significance is lower for these variables compared to the strong drop in output on impact. The cumulative effect on output is similar between high and low FX debt countries. In online appendix A5, we use time-varying variables for IAPOC index and FX debt, getting similar results.

IV. The Recent Episode: 2022–2023 Fed Hikes

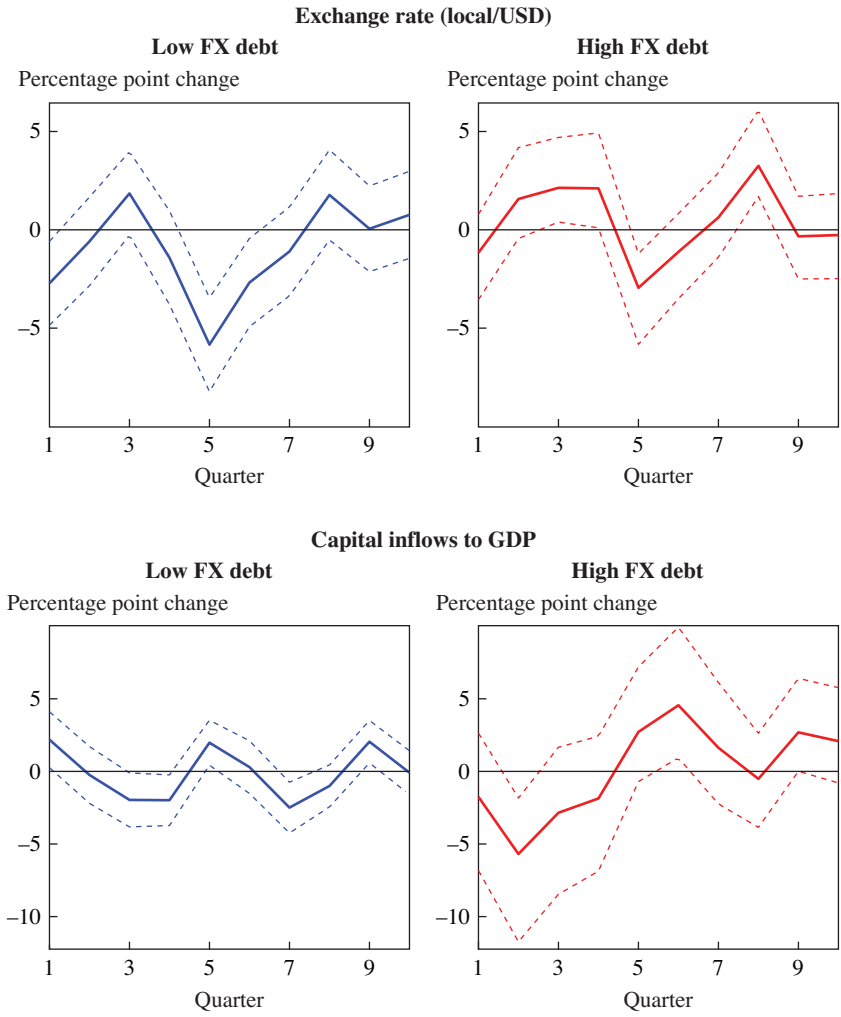
"Resilience" is the buzz word for 2022–2023. While it is often used in the context of the US economy, which has avoided a recession despite experiencing the steepest interest rate hikes in decades, the story of emerging

Figure 15. International Transmission of the Fed Hikes: The Role of Balance Sheet FX Vulnerabilities with Global Controls (Gertler-Karadi Surprises)



(*continued on next page*)

Figure 15. International Transmission of the Fed Hikes: The Role of Balance Sheet FX Vulnerabilities with Global Controls (Gertler-Karadi Surprises) (*Continued*)



Source: Authors' calculations.

Note: Impulse responses of the twelve-month US Treasury rate, instrumented by monthly weighted raw surprises in three-month federal funds futures from Gertler and Karadi (2015), are obtained from panel local projections. Confidence intervals at 90 percent (calculated using Newey-West standard errors) are shown by the dashed lines. Controls include dollar shock, average oil price index, and median trade balance, and four lags of the dependent variable, twelve-month US Treasury rate, output growth and inflation differentials with the United States, and the instrument. In this case, we did not add four lags of dollar shock, average oil price index, and median trade balance because of the limited sample. Global controls enter contemporaneously. Dependent variables include real GDP in logs, CPI in logs, quarter-to-quarter nominal exchange rate growth (domestic currency/US dollar), and capital inflows to GDP ratio. See text for the definitions of high and low FX debt countries.

markets is even more remarkable. Projections for global growth in 2023 are primarily fueled by emerging markets, and impressively, the top twenty-five emerging markets all surpassed their 2022 forecasts (IMF 2023).

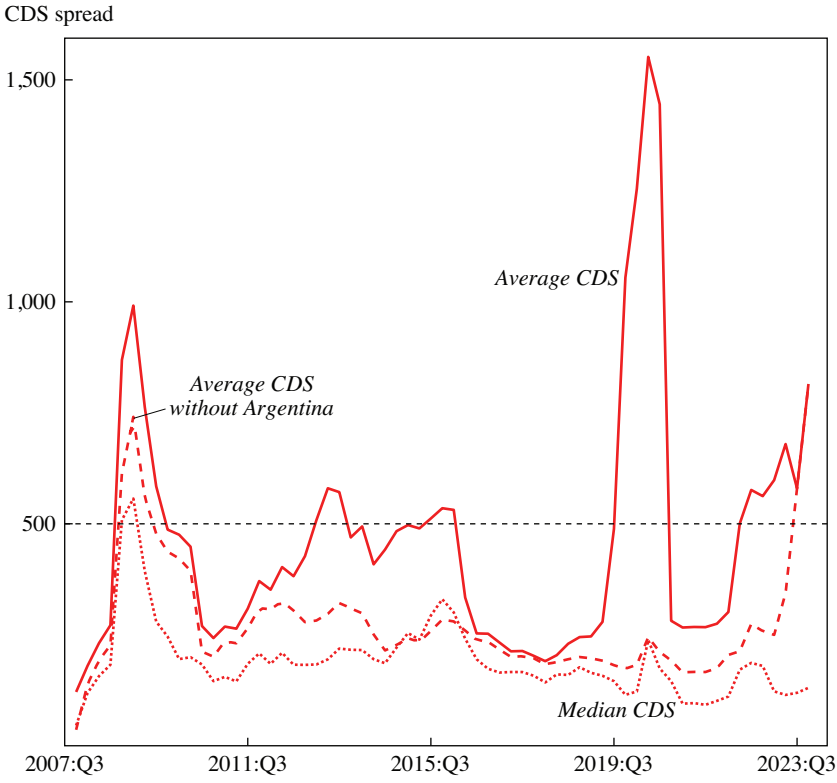
As is widely acknowledged, and as we confirm in this paper, rising US interest rates historically created challenges for emerging markets. This time is different as most emerging markets managed to establish monetary and financial discipline, marked by credible monetary policies and reduced FX debt, as shown in figures 1 and 2 respectively. In the recent period, they began raising rates ahead of advanced economies as soon as the COVID-19 inflation hit their economies. This shows improved monetary policy credibility since the monetary policy is responding to their own inflation rather than to the US policy or the exchange rate developments. Their statements were clear on why they were raising interest rates: not to mimic the US policy for currency defense, but rather to re-anchor the rising inflation expectations (Carvalho and Nechio 2023).

The first piece of evidence for this time being different is that the main risk spread—the credit default swaps (CDS)—did not move at all for emerging markets, as shown in figure 16. Compared to 2008 when the CDS spreads spiked for both average and median emerging markets, this time around they actually went down for the median emerging market. For the average emerging market, there was a huge spike totally driven by Argentina in 2020 when the pandemic started. In 2022 when the Federal Reserve started hiking, the median emerging market spread went down and the average emerging market spread (without Argentina) went up very little, less than what happened in the taper tantrum. The CDS spread captures the default risk of governments on dollar-denominated bonds. Clearly this risk was very low.

Figure 17 shows, relative to the first quarter of 2022, the change in the twelve-month UIP deviations for advanced economies and emerging markets. Investigating UIP spread on top of the CDS spread is useful since the UIP risk spread captures the risk premium due to currency depreciations and passes through the domestic lending rates one to one. Relative to our findings in previous sections, changes in the UIP premia are much smaller for emerging markets than advanced economies. Consistently, figure 18 shows similar exchange rate movements in advanced economies and emerging markets and in high and low credibility countries. This is because there is not much difference now between these countries given the improvement in monetary policy credibility, where the low value is 0.51 and the high value is 0.6.

We do not have enough observations to run local projections with the US monetary policy shocks starting in 2022:Q1. We have run an alternative

Figure 16. Credit Default Swaps (CDS) in the Recent Episode



Source: Refinitiv Datastream.

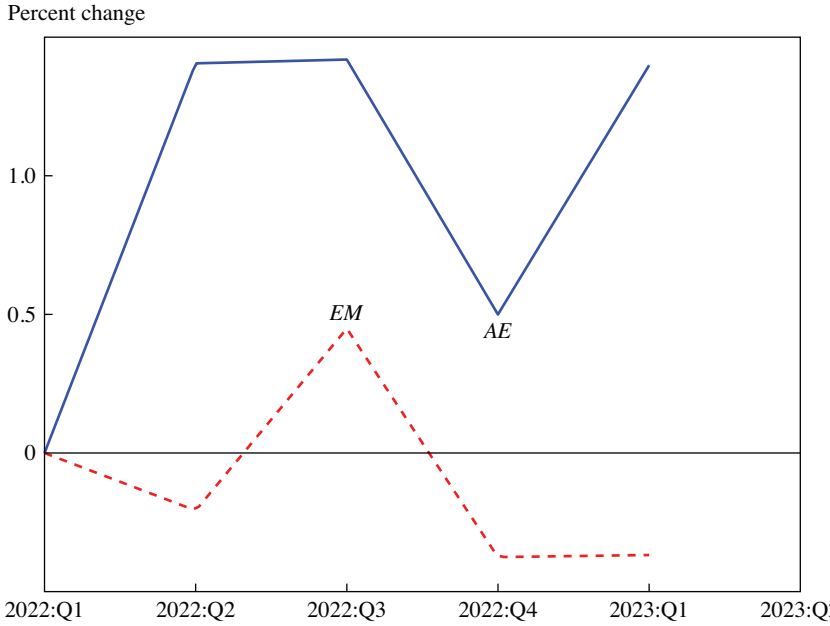
Note: CDS for fifteen emerging markets: Argentina, Egypt, Guatemala, India, Kazakhstan, Korea, Malta, Mexico, Morocco, Pakistan, Serbia, Singapore, Slovak Republic, Thailand, and Uruguay.

panel regression to nail down this point that emerging markets became resilient to sudden stops related to Fed hikes, as follows:

$$y_{ct} = \alpha_c + \delta_{year} + \gamma_1 Q_1 + \gamma_2 Q_2 + \gamma_3 Q_3 + \gamma_4 Q_4 + \varepsilon_{ct}, \tag{7}$$

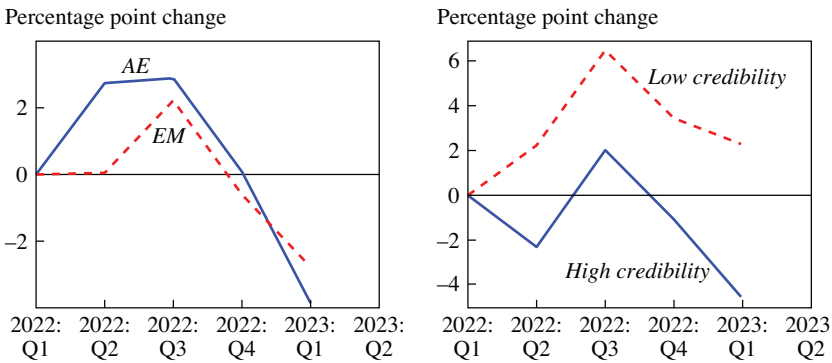
where y_{ct} is the dependent variable and includes exchange rate depreciation (year-on-year), real GDP growth (year-to-year), real investment growth (year-to-year), and trade balance/GDP. All variables are in percentages. Controls include country fixed effects (α_c), year fixed effects (δ_{year}), and four dummies. The first dummy takes the value one when quarter zero is the sudden stop and so on ($\{Q_i\}_{i=1}^{4}$). We run equation (7) in two recent time

Figure 17. UIP during the 2022–2023 Fed Hikes



Source: IMF International Financial Statistics; Consensus Economics; and authors' calculations.
Note: This figure shows the percentage change in the twelve-month UIP deviations relative to 2022:Q1 for advanced economies (AEs) and emerging markets (EMs). UIP deviations are calculated as explained in the data section.

Figure 18. Exchange Rates during the 2022–2023 Fed Hikes



Source: IMF International Financial Statistics; Consensus Economics; and authors' calculations.
Note: The growth rate of nominal exchange rate (domestic currency/US dollar) with respect to 2022:Q1.

periods in panels B and C of table 3 and show historical results for the same regression in panel A from Eichengreen and Gupta (2017). Panel A covers forty-six sudden stops during the period 1991–2015 for twenty emerging markets in 1991, twenty-eight in 1995, and thirty-four from 2000 onward. Panel B covers the only sudden stop in March 2020 for our emerging markets. Panel C covers the Federal Reserve's signal of hikes as of December 2021, also for our emerging markets. Panels B and C don't include year fixed effects.

As table 3 clearly shows, the sudden stops of March 2020 and the Federal Reserve signaling a hike in December 2021 markedly differ from previous sudden stop episodes. Notably, there was a much lower currency depreciation, a less persistent drop in GDP and investment, and negligible impact on the trade balance. Historically, sudden stops are linked with current account reversals, which are typically evident by the third quarter. However, even in the fourth quarter following the Federal Reserve's rate hike signal, while there was a reversal, it did not significantly affect investment, indicating a newfound resilience to such shocks, which may plausibly be ascribed to enhanced monetary policy credibility and reduced foreign exchange debt.

V. Conclusion

We ask why emerging markets showed resilience in the face of sharp and quick Fed hikes during the last two years. In the 1980s and 1990s, the global transmission of Fed hikes rooted in financial channels, often resulted in adverse repercussions for emerging markets characterized by sudden stops, increased UIP premia, capital outflows, and sharp recessions. In the post-COVID-19 era, however, none of these events were observed. We argue that this is due to the improved monetary policy credibility and lower dollar-denominated debt in emerging markets this time around compared to historical episodes.

With diminished risk sensitivity and reduced volatility of capital flows, emerging markets seem to be better insulated against the shifts in global investor sentiment and the risk-aversion shocks, which are associated with the Fed hikes. During the last two years, despite the sharply rising US interest rates, emerging market spreads have stayed stable with no major financial crises. Although inflation also rose quite dramatically in emerging markets, inflation expectations have remained largely anchored thanks to their improved monetary policy credibility.

Table 3. Sudden Stops in Emerging Markets

	(1) *ER depreciation*	*(2)* *GDP growth (yoy)*	*(3)* *Investment growth (yoy)*	*(4)* *Trade balance/ GDP*
Panel A: 1991–2015 (46 sudden stops)				
Quarter 1	10.126***	–2.270***	–6.019**	–0.662
	(4.37)	(3.09)	(2.75)	(1.12)
Quarter 2	12.853***	–5.521***	–9.038**	1.045
	(3.40)	(4.97)	(2.17)	(1.14)
Quarter 3	3.514**	–5.845***	–16.643***	2.506*
	(2.39)	(4.51)	(3.83)	(2.32)
Quarter 4	5.621	–5.193***	–14.447**	3.272***
	(1.67)	(2.95)	(2.46)	(2.84)
N	2,658	2,236	2,031	2,076
Adjusted R^2	0.027	0.07	0.03	0.01
Panel B: 2020–2021 (sudden stop of March 2020)				
Quarter 1	3.389***	–11.478***	–19.971***	–1.084
	(3.59)	(8.62)	(5.05)	(1.18)
Quarter 2	–3.608***	–3.702***	–6.291	0.618
	(3.82)	(2.74)	(1.59)	(0.67)
Quarter 3	–2.941***	–1.124	–0.693	–1.412
	(3.11)	(0.83)	(0.18)	(1.53)
Quarter 4	–3.361***	2.053	5.554	–1.142
	(3.56)	(1.52)	(1.40)	(1.24)
N	130	127	110	120
Adjusted R^2	0.463	0.549	0.409	–0.131
Panel C: 2021–2022 (Federal Reserve signal of 2020 hikes of December 2021)				
Quarter 1	–0.643	–0.286	–0.521	0.537
	(0.44)	(0.44)	(0.37)	(0.59)
Quarter 2	–1.271	–1.355**	0.339	0.914
	(0.86)	(2.06)	(0.24)	(1.00)
Quarter 3	2.201	–1.406**	0.778	–0.281
	(1.50)	(2.08)	(0.52)	(0.30)
Quarter 4	–0.506	–3.135***	–0.307	2.890***
	(0.34)	(4.64)	(0.2)	(2.84)
N	130	121	104	107
Adjusted R^2	0.258	0.567	0.371	–0.086

Source: Panel A is reproduced from Eichengreen and Gupta (2017), copyright *Economía Chilena*; panels B and C are based on authors' calculations.

Note: This table summarizes the panel regression estimates of $y_{ct} = \alpha_c + \delta_{year} + \sum_{k=1}^{4} \gamma_k Q_k + \varepsilon_{it}$, where y_{ct} is the outcome for country c in quarter t, and α and δ are country and year fixed effects. Panels B and C don't include the year fixed effects. Q_k is a dummy variable that takes the value of one when t is k quarters after the sudden stop period. Dependent variables include exchange rate depreciation, real GDP growth (year-to-year), real investment growth (year-to-year), and trade balance/GDP. All variables are in percentages; t statistics are in parentheses. Panel A covers sudden stops for twenty emerging markets (EMs) in 1991, twenty-eight in 1995, and thirty-four from 2000 onward. Panel B covers the sudden stop in March 2020 for the EMs studied in this analysis (summarized in table 1). Panel C covers the Federal Reserve's signal of 2020 hikes in December 2021, also for the EMs studied in this paper. Data are quarterly.

Coefficient level of significance: *10 percent, **5 percent, ***1 percent.

ACKNOWLEDGMENTS The authors thank Jose Ignacio Cristi Le-Fort and Mariana Sans for their outstanding research assistance, and Hudson Hinshaw and Omer Faruk Akbal for their superb help with the data. The views expressed and arguments employed in this paper are solely those of the authors and do not necessarily reflect the official views of the International Monetary Fund and Organisation for Economic Co-operation and Development or its member countries. Any errors or omissions are the responsibility of the authors.

References

Aghion, Philippe, Philippe Bacchetta, and Abhijit Banerjee. 2001. "Currency Crises and Monetary Policy in an Economy with Credit Constraints." *European Economic Review* 45, no. 7: 1121–50.

Aguiar, Mark. 2005. "Investment, Devaluation, and Foreign Currency Exposure: The Case of Mexico." *Journal of Development Economics* 78, no. 1: 95–113.

Akinci, Özge, Şebnem Kalemli-Özcan, and Albert Queralto. 2021. "Uncertainty Shocks, Capital Flows, and International Risk Spillovers." Working Paper 30026. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30026.

Akinci, Özge, and Albert Queralto. 2023. "Exchange Rate Dynamics and Monetary Spillovers with Imperfect Financial Markets." *Review of Financial Studies* 37, no. 2: 309–55.

Alesina, Alberto, and Lawrence H. Summers. 1993. "Central Bank Independence and Macroeconomic Performance: Some Comparative Evidence." *Journal of Money, Credit and Banking* 25, no. 2: 151–62.

Alfaro, Laura, Mauricio Calani, and Liliana Varela. 2023. "Granular Corporate Hedging under Dominant Currency." Working Paper 28910. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w28910.

Alfaro, Laura, Şebnem Kalemli-Özcan, and Vadym Volosovych. 2014. "Sovereigns, Upstream Capital Flows, and Global Imbalances." *Journal of the European Economic Association* 12, no. 5: 1240–84.

Avdjiev, Stefan, Wenxin Du, Catherine Koch, and Hyun Song Shin. 2019. "The Dollar, Bank Leverage, and Deviations from Covered Interest Parity." *American Economic Review: Insights* 1, no. 2: 193–208.

Avdjiev, Stefan, Bryan Hardy, Şebnem Kalemli-Özcan, and Luis Servén. 2022. "Gross Capital Flows by Banks, Corporates, and Sovereigns." *Journal of the European Economic Association* 20, no. 5: 2098–135.

Basu, Suman S., Emine Boz, Gita Gopinath, Francisco Roch, and Filiz D. Unsal. 2020. "A Conceptual Model for the Integrated Policy Framework." Working Paper 2020/121. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2020/07/07/A-Conceptual-Model-for-the-Integrated-Policy-Framework-49558.

Bauer, Michael D., Ben S. Bernanke, and Eric Milstein. 2023. "Risk Appetite and the Risk-Taking Channel of Monetary Policy." *Journal of Economic Perspectives* 37, no. 1: 77–100.

Bauer, Michael D., and Eric T. Swanson. 2023. "A Reassessment of Monetary Policy Surprises and High-Frequency Identification." *NBER Macroeconomics Annual* 37, no. 1: 87–155.

Bekaert, Geert, Marie Hoerova, and Marco Lo Duca. 2013. "Risk, Uncertainty and Monetary Policy." *Journal of Monetary Economics* 60, no. 7: 771–88.

Bems, Rudolfs, Francesca Caselli, Francesco Grigoli, and Bertrand Gruss. 2021. "Expectations' Anchoring and Inflation Persistence." *Journal of International Economics* 132:103516.

Bénétrix, Agustín, Deepali Gautam, Luciana Juvenal, and Martin Schmitz. 2019. "Cross-Border Currency Exposures." Working Paper 2019/299. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2019/12/27/Cross-Border-Currency-Exposures-48876.

Bianchi, Javier, Saki Bigio, and Charles Engel. 2021. "Scrambling for Dollars: International Liquidity, Banks and Exchange Rates." Working Paper 29457. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w29457.

Bianchi, Javier, and Guido Lorenzoni. 2022. "The Prudential Use of Capital Controls and Foreign Currency Reserves." In *Handbook of International Economics: International Macroeconomics, Volume 6*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Amsterdam: North-Holland.

Boehm, Christoph E., and T. Niklas Kroner. 2023. "The US, Economic News, and the Global Financial Cycle." Working Paper 30994. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30994.

Bruno, Valentina, and Hyun Song Shin. 2015. "Capital Flows and the Risk-Taking Channel of Monetary Policy." *Journal of Monetary Economics* 71:119–32.

Burstein, Ariel, and Gita Gopinath. 2014. "International Prices and Exchange Rates." In *Handbook of International Economics, Volume 4*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Amsterdam: North-Holland.

Calvo, Guillermo A., Leonardo Leiderman, and Carmen M. Reinhart. 1993. "Capital Inflows and Real Exchange Rate Appreciation in Latin America: The Role of External Factors." *IMF Staff Papers* 40, no. 1: 108–51.

Calvo, Guillermo A., Leonardo Leiderman, and Carmen M. Reinhart. 1996. "Inflows of Capital to Developing Countries in the 1990s." *Journal of Economic Perspectives* 10, no. 2: 123–39.

Calvo, Guillermo A., and Carmen M. Reinhart. 2002. "Fear of Floating." *Quarterly Journal of Economics* 117, no. 2: 379–408.

Carvalho, Carlos, and Fernanda Nechio. 2023. "Challenges to Disinflation: The Brazilian Experience." *Brookings Papers on Economic Activity*, Spring, 217–41.

Céspedes, Luis Felipe, Roberto Chang, and Andrés Velasco. 2004. "Balance Sheets and Exchange Rate Policy." *American Economic Review* 94, no. 4: 1183–93.

Chari, Anusha, Karlye Dilts Stedman, and Christian Lundblad. 2020. "Capital Flows in Risky Times: Risk-On/Risk-Off and Emerging Market Tail Risk." Working Paper 27927. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w27927.

Chari, Anusha, Karlye Dilts Stedman, and Christian Lundblad. 2021. "Taper Tantrums: Quantitative Easing, Its Aftermath, and Emerging Market Capital Flows." *Review of Financial Studies* 34, no. 3: 1445–508.

Cieslak, Anna, and Andreas Schrimpf. 2019. "Non-Monetary News in Central Bank Communication." *Journal of International Economics* 118: 293–315.

Cook, David. 2004. "Monetary Policy in Emerging Markets: Can Liability Dollarization Explain Contractionary Devaluations?" *Journal of Monetary Economics* 51, no. 6: 1155–81.

Das, Mitali, Şebnem Kalemli-Özcan, Damien Puy, and Liliana Varela. 2020. "Emerging Markets' Hidden Debt Risk." Project Syndicate, May 20. https://www.project-syndicate.org/commentary/covid19-emerging-market-firms-foreign-currency-debt-risk-by-mitali-das-et-al-2020-05?barrier=accesspaylog.

Dedola, Luca, Giulia Rivolta, and Livio Stracca. 2017. "If the Fed Sneezes, Who Catches a Cold?" *Journal of International Economics* 108, no. S1: S23–S41.

Degasperi, Riccardo, Seokki Hong, and Giovanni Ricco. 2023. "The Global Transmission of US Monetary Policy." Working Paper 2023-02. Palaiseau: Center for Research in Economics and Statistics. https://crest.science/wp-content/uploads/2023/01/2023-02.pdf.

De Leo, Pierre, Gita Gopinath, and Şebnem Kalemli-Özcan. 2022. "Monetary Policy Cyclicality in Emerging Economies." Working Paper 30458. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30458.

Devereux, Michael B., Charles Engel, and Steve Pak Yeung Wu. 2023. "Collateral Advantage: Exchange Rates, Capital Flows and Global Cycles." Working Paper 31164. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31164.

Diaz-Alejandro, Carlos F. 1983. "Stories of the 1930s for the 1980s." In *Financial Policies and the World Capital Market: The Problem of Latin American Countries*, edited by Pedro Aspe Armella, Rudiger Dornbusch, and Maurice Obstfeld. Chicago: University of Chicago Press.

Di Giovanni, Julian, Şebnem Kalemli-Özcan, Mehmet Fatih Ulu, and Yusuf Soner Baskaya. 2022. "International Spillovers and Local Credit Cycles." *Review of Economic Studies* 89, no. 2: 733–73.

Di Giovanni, Julian, and John Rogers. 2023. "The Impact of US Monetary Policy on Foreign Firms." *IMF Economic Review*. https://link.springer.com/article/10.1057/s41308-023-00218-7.

Dincer, N. Nergiz, and Barry Eichengreen. 2014. "Central Bank Transparency and Independence: Updates and New Measures." *International Journal of Central Banking* 10, no. 1: 189–259.

Du, Wenxin, and Jesse Schreger. 2016. "Local Currency Sovereign Risk." *Journal of Finance* 71, no. 3: 1027–70.

Economist. 2020. "Turkey's Bizarre Economic Experiment Enters A New Phase." June 1. https://www.economist.com/finance-and-economics/2023/06/01/turkeys-bizarre-economic-experiment-enters-a-new-phase.

Eichengreen, Barry, and Poonam Gupta. 2017. "Managing Sudden Stops." *Economía Chilena* 20, no. 2: 4–41.

Eichengreen, Barry, and Ricardo Hausmann. 1999. "Exchange Rates and Financial Fragility." In *Economic Policy Symposium Proceedings: New Challenges for Monetary Policy*. Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Eichengreen, Barry, Ricardo Hausmann, and Ugo Panizza. 2005. "The Pain of Original Sin." In *Other People's Money: Debt Denomination and Financial Instability in Emerging Market Economies*, edited by Barry Eichengreen and Ricardo Hausmann. Chicago: University of Chicago Press.

Eichengreen, Barry, and Richard Portes. 1987. "The Anatomy of Financial Crises." In *Threats to International Financial Stability*, edited by Richard Portes and Alexander K. Swoboda. Cambridge: Cambridge University Press.

Fan, Jingting, and Şebnem Kalemli-Özcan. 2016. "Emergence of Asia: Reforms, Corporate Savings, and Global Imbalances." *IMF Economic Review* 64, no. 2: 239–67.

Forbes, Kristin J., Ida Hjortsoe, and Tsvetelina Nenova. 2018. "The Shocks Matter: Improving Our Estimates of Exchange Rate Pass-Through." *Journal of International Economics* 114: 255–75.

Forbes, Kristin J., and Francis E. Warnock. 2012. "Capital Flow Waves: Surges, Stops, Flight, and Retrenchment." *Journal of International Economics* 88, no. 2: 235–51.

Fukui, Masao, Emi Nakamura, and Jón Steinsson. 2023. "The Macroeconomic Consequences of Exchange Rate Depreciations." Working Paper 31279. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31279.

Gabaix, Xavier, and Matteo Maggiori. 2015. "International Liquidity and Exchange Rate Dynamics." *Quarterly Journal of Economics* 130, no. 3: 1369–420.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics* 7, no. 1: 44–76.

Gertler, Mark, and Nobuhiro Kiyotaki. 2010. "Financial Intermediation and Credit Policy in Business Cycle Analysis." In *Handbook of Monetary Economics, Volume 3*, edited by Benjamin M. Friedman and Michael Woodford. Amsterdam: North-Holland.

Gopinath, Gita. 2016. "The International Price System." In *Economic Policy Symposium Proceedings: Designing Resilient Monetary Policy Frameworks for the Future.* Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Gopinath, Gita, and Brent Neiman. 2014. "Trade Adjustment and Productivity in Large Crises." *American Economic Review* 104, no. 3: 793–831.

Gourinchas, Pierre-Oliver. 2018. "Monetary Policy Transmission in Emerging Markets: An Application to Chile." In *Monetary Policy and Global Spillovers: Mechanisms, Effects and Policy Measures*, edited by Enrique G. Mendoza, Ernesto Pastén, and Diego Saravia. Santiago: Central Bank of Chile.

Gourinchas, Pierre-Olivier, and Hélène Rey. 2022. "Exorbitant Privilege and Exorbitant Duty." Discussion Paper DP16944. London: Centre for Economic Policy Research.

Gürkaynak, Refet S., Brian P. Sack, and Eric T. Swanson. 2004. "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements." Finance and Economics Discussion Series. Washington: Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/pubs/feds/2004/200466/200466pap.pdf.

Hofmann, Boris, Nikhil Patel, and Steve Pak Yeung Wu. 2022. "Original Sin Redux: A Model-Based Evaluation." Working Paper 1004. Basel: Bank for International Settlements. https://www.bis.org/publ/work1004.htm.

Iacoviello, Matteo, and Gaston Navarro. 2019. "Foreign Effects of Higher U.S. Interest Rates." *Journal of International Money and Finance* 95: 232–50.

Ilzetzki, Ethan, Carmen M. Reinhart, and Kenneth S. Rogoff. 2022. "Rethinking Exchange Rate Regimes." In *Handbook of International Economics: International Macroeconomics, Volume 6*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Amsterdam: North-Holland.

International Monetary Fund (IMF). 2022. "Review of the Institutional View on the Liberalization and Management of Capital Flows." Policy Paper 2022/008. Washington: Author.

International Monetary Fund (IMF). 2023. *World Economic Outlook, April 2023: A Rocky Recovery*. Washington: Author.

Itskhoki, Oleg, and Dmitry Mukhin. 2022. "Sanctions and the Exchange Rate." Working Paper 30009. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30009.

Jarociński, Marek, and Peter Karadi. 2020. "Deconstructing Monetary Policy Surprises—The Role of Information Shocks." *American Economic Journal: Macroeconomics* 12, no. 2: 1–43.

Jiang, Erica Xuewei, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2023. "Monetary Tightening and U.S. Bank Fragility in 2023: Mark-to-Market Losses and Uninsured Depositor Runs?" Working Paper 31048. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31048.

Jiang, Zhengyang, Arvind Krishnamurthy, and Hanno Lustig. 2021. "Foreign Safe Asset Demand and the Dollar Exchange Rate." *Journal of Finance* 76, no. 3: 1049–89.

Jordà, Òscar. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95, no. 1: 161–82.

Kalemli-Özcan, Şebnem. 2019. "U.S. Monetary Policy and International Risk Spillovers." In *Economic Policy Symposium Proceedings: Challenges for Monetary Policy.* Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Kalemli-Özcan, Şebnem. 2022. "Turkey's Risky Inflation Experiment." Project Syndicate. January 4. https://www.project-syndicate.org/commentary/flawed-economic-theory-driving-turkey-inflation-by-sebnem-kalemli-ozcan-2022-01?barrier=accesspaylog.

Kalemli-Özcan, Şebnem, Herman Kamil, and Carolina Villegas-Sanchez. 2016. "What Hinders Investment in the Aftermath of Financial Crises: Insolvent Firms or Illiquid Banks?" *Review of Economics and Statistics* 98, no. 4: 756–69.

Kalemli-Özcan, Şebnem, Xiaoxi Liu, and Ilhyock Shim. 2021. "Exchange Rate Fluctuations and Firm Leverage." *IMF Economic Review* 69, no. 1: 90–121.

Kalemli-Özcan, Şebnem, and Liliana Varela. 2021. "Five Facts about the UIP Premium." Working Paper 28923. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w28923.

Kamil, Herman. 2004. "A New Database on the Currency Composition and Maturity Structure of Firms' Balance Sheets in Latin America, 1990–2002." Washington: Inter-American Development Bank.

Kekre, Rohan, and Moritz Lenel. 2021. "The Flight to Safety and International Risk Sharing." Working Paper 29238. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w29238.

Kilian, Lutz, and Helmut Lütkepohl. 2017. *Structural Vector Autoregressive Analysis*. Cambridge: Cambridge University Press.

Krugman, Paul. 1999. "Balance Sheets, the Transfer Problem, and Financial Crises." *International Tax and Public Finance* 6: 459–72.

Kuttner, Kenneth N. 2001. "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market." *Journal of Monetary Economics* 47, no. 3: 523–44.

McCauley, Robert N., Patrick McGuire, and Vladyslav Sushko. 2015. "Dollar Credit to Emerging Market Economies." *BIS Quarterly Review*, December.

Mendoza, Enrique G., and Vivian Z. Yue. 2012. "A General Equilibrium Model of Sovereign Default and Business Cycles." *Quarterly Journal of Economics* 127, no. 2: 889–946.

Miranda-Agrippino, Silvia, and Hélène Rey. 2020. "U.S. Monetary Policy and the Global Financial Cycle." *Review of Economic Studies* 87, no. 6: 2754–76.

Miranda-Agrippino, Silvia, and Giovanni Ricco. 2023. "Identification with External Instruments in Structural VARs." *Journal of Monetary Economics* 135: 1–19.

Nakamura, Emi, and Jón Steinsson. 2018. "High-Frequency Identification of Monetary Non-neutrality: The Information Effect." *Quarterly Journal of Economics* 133, no. 3: 1283–330.

Obstfeld, Maurice. 2015. "Trilemmas and Tradeoffs: Living with Financial Globalization." In *Global Liquidity, Spillovers to Emerging Markets and Policy Responses*, edited by Claudio E. Raddatz, Diego Saravia, and Jaume Ventura. Santiago: Central Bank of Chile.

Obstfeld, Maurice, and Haonan Zhou. 2022. "The Global Dollar Cycle." *Brookings Papers on Economic Activity*, Fall, 361–427.

Plagborg-Møller, Mikkel, and Christian K. Wolf. 2021. "Local Projections and VARs Estimate the Same Impulse Responses." *Econometrica* 89, no. 2: 955–80.

Reinhart, Carmen M. 2000. "The Mirage of Floating Exchange Rates." *American Economic Review* 90, no. 2: 65–70.

Reinhart, Carmen M., and Vincent R. Reinhart. 2009. "Capital Flow Bonanzas: An Encompassing View of the Past and Present." In *NBER International Seminar on Macroeconomics 2008*, edited by Jeffrey Frankel and Christopher Pissarides. Chicago: University of Chicago Press.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, N.J.: Princeton University Press.

Rey, Hélène. 2013. "Dilemma not Trilemma: The Global Financial Cycle and Monetary Policy Independence." In *Economic Policy Symposium Proceedings: Global Dimensions of Unconventional Monetary Policy.* Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Rogers, John H., Bo Sun, and Wenbin Wu. 2023. "Drivers of the Global Financial Cycle." Working Paper. Social Science Research Network, March 22. https://ssrn.com/abstract=4397119.

Rogoff, Kenneth. 1985. "The Optimal Degree of Commitment to an Intermediate Monetary Target." *Quarterly Journal of Economics* 100, no. 4: 1169–89.

Rogoff, Kenneth. 2023. "The Stunning Resilience of Emerging Markets." Project Syndicate, October 31. https://www.project-syndicate.org/commentary/macroeconomic-orthodoxy-saved-emerging-markets-from-debt-crisis-by-kenneth-rogoff-2023-10.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual* 4:121–70.

Salomao, Juliana, and Liliana Varela. 2022. "Exchange Rate Exposure and Firm Dynamics." *Review of Economic Studies* 89, no. 1: 481–514.

Sargent, Thomas J., and Neil Wallace. 1981. "Some Unpleasant Monetarist Arithmetic." *Federal Reserve Bank of Minneapolis Quarterly Review* 5, no. 3: 1–17.

Schneider, Martin, and Aaron Tornell. 2004. "Balance Sheet Effects, Bailout Guarantees and Financial Crises." *Review of Economic Studies* 71, no. 3: 883–913.

Stock, James H., and Mark W. Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *Economic Journal* 128, no. 610: 917–48.

Tankersley, Jim, Ana Swanson, and Matt Phillips. 2018. "Trump Hits Turkey When It's Down, Doubling Tariffs." *New York Times*, August 10. https://www.nytimes.com/2018/08/10/us/politics/trump-turkey-tariffs-currency.html.

Unsal, Filiz D., Chris Papageorgiou, and Hendre Garbers. 2022. "Monetary Policy Frameworks: An Index and New Evidence." Working Paper 2022/022. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/01/28/Monetary-Policy-Frameworks-An-Index-and-New-Evidence-512228.

Comments and Discussion

COMMENT BY

KRISTIN FORBES Kalemli-Özcan and Unsal ask an important question: why have many emerging markets been so resilient to the sharp tightening in US monetary policy over 2022–2023? The authors propose two answers: increased monetary policy credibility and lower levels of FX-denominated debt. This topic is timely and provides insights on what policies emerging markets should prioritize to reduce their vulnerability in the future.[1]

I will divide my comments into three parts: a quick summary of the main sections of the paper (with a few editorial comments), the broader context of how emerging market resilience has changed over a longer period, and some concerns about the data and omitted variables.

QUICK PAPER SUMMARY This paper covers a lot of ground. It begins with an overview of recent literature on how US monetary policy is transmitted to other countries. It focuses on financial channels of transmission, such as through risk premia, the cost of capital, and exchange rate effects. This discussion helps motivate the choice of variables included later in the empirical analysis. It also includes a case study of the impact of US interest rates on Canada and Mexico—which is a useful example to make the channels concrete. It also provides a description of the key variables for Turkey—an example that highlights the challenges of sorting out the multiple

1. It is worth noting that the authors' focus on the recent resilience of emerging markets describes many middle-income emerging markets, but not all. A number of emerging markets, and many developing economies, are struggling with slow growth, high inflation, and an inability to repay debt—problems aggravated by the recent increases in global interest rates. Other major emerging markets (such as Argentina) are on the verge of default. These situations—and particularly the current challenges of highly indebted, low-income countries—are not the focus of the paper.

Brookings Papers on Economic Activity, Fall 2023: 226–248

interrelationships between the key variables of interest in this paper.[2] This motivational section does not discuss the role of commodity markets and argues that the trade channel of transmission was not important during this period. While much of the recent literature (including that cited in the paper) highlights how shocks that affect the exchange rate may not generate the standard Mundell-Fleming effects through trade, I worry that ignoring commodity markets and trade may miss factors that were important during the 2021–2023 episode that motivates this paper. For example, early in this period, commodity prices spiked as countries reopened and after the invasion of Ukraine, boosting revenues, FX earnings, and exchange rates for many of the commodity-exporting emerging markets that are central to the analysis. These changes in commodity prices—which have heterogenous effects in different countries and therefore cannot be captured in time dummies—could be an important factor contributing to the recent resilience in many emerging markets in the sample.

In the next section of their paper, Kalemli-Özcan and Unsal discuss the main data sources used for the empirical analysis, highlighting a new measure of monetary policy credibility from Unsal, Papageorgiou, and Garbers (2022) and a measure of FX exposure from Fan and Kalemli-Özcan (2016) and Kalemli-Özcan, Liu, and Shim (2021). These variables are central to the paper and could be an important contribution to the literature—especially the measure of monetary policy credibility. I will discuss these data sources in more detail below, but I hope that the authors will be able to share these data in the future as they could be an important resource.

The authors then estimate their baseline model of the impact of US monetary tightening using a local projections method. They focus on the impact of the "surprise" component of US monetary policy—and since there are several different approaches to estimating this—provide an extensive set of sensitivity tests using different proxies for monetary policy shocks. They estimate the impact of these shocks on GDP, the exchange rate, CPI, UIP deviations, and capital inflows (although they only report a subset of results for each main test), and focus on the impact based on three different country characteristics: advanced economies versus emerging markets, emerging markets with more and less central bank credibility, and emerging markets with more and less FX exposure. Many of the results

2. More specifically, Turkey has recently experienced very high inflation and a sharp currency depreciation, combined with high FX debt and a large improvement in policy credibility since 2007 (which incorporates a sizable improvement from 2007 to 2018 combined with a small deterioration from 2018 to 2021).

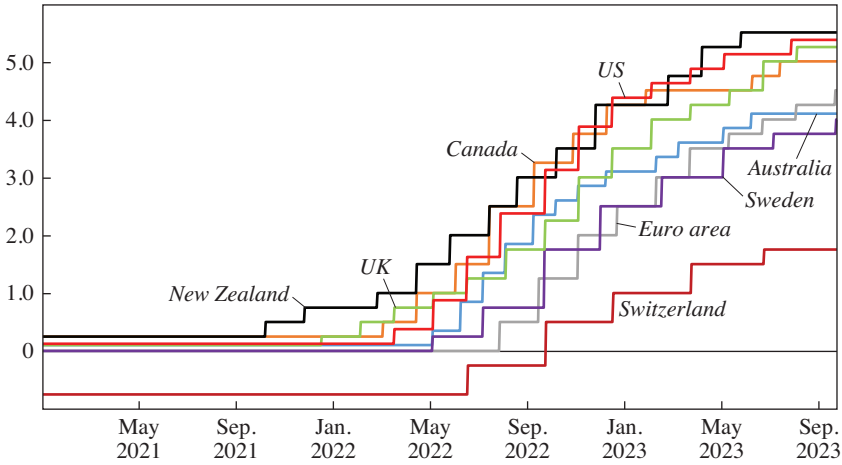
move in the expected direction and support the arguments outlined at the start of the paper—particularly a more negative impact on GDP and the UIP premium in emerging markets and countries with weaker central bank credibility. Some of the results, however, show some odd patterns and are not what I expected—such as the patterns for capital inflows and relative resilience of countries with more FX debt (particularly for GDP).

The last section of the paper contains the punchline: do improvements in central bank credibility and reduced FX exposure explain the recent resilience of emerging markets to the rapid tightening in US monetary policy? Unfortunately this section of the paper cannot yet be completed given the lags in obtaining key data and the fact that not enough time has passed to use the authors' methodology. The authors are aware of these limitations and show some regression results using a different framework that confirms emerging markets have been more resilient during this period (based on criteria such as their exchange rates, GDP growth, investment growth, and the trade balance). Unfortunately, they are not able to test the key hypothesis of the paper: did this resilience result from improved central bank credibility and lower levels of FX debt? I hope the authors will return to this analysis in the future when additional quarters of data are available to extend their analysis for this important case study.

IMPROVED RESILIENCE IN EMERGING MARKETS: THE BROADER CONTEXT The paper is motivated by the question of why many emerging markets have been fairly resilient to the sharp tightening in US monetary policy over 2022–2023. This is a timely and important question. Figure 1 shows that the United States is not the only major economy to raise its policy interest rate sharply—and this does not even incorporate the other ways in which monetary policy has been tightened (such as through unwinding central bank asset holdings). The tightening in monetary policy has been widespread and has occurred much faster and with rates increasing to a much higher level than forecasters were expecting at earlier stages in this cycle. For example, on January 1, 2021, the US terminal rate (i.e., the peak of the policy rate during this tightening cycle) was expected to be 85 basis points; on January 1, 2022, it was expected to be 1.72 percent; and at the start of June 2022 (even after the Federal Reserve had raised its policy rate by 50 basis points in one meeting), the expected terminal rate was only 3.0 percent.[3] This is well below the current band for the federal funds rate of 5.25–5.50 percent (in December 2023)—highlighting how much of this tightening in US monetary policy

3. The terminal rate data are from Morgan Stanley and available at Bloomberg, "Rates and Bonds," https://www.bloomberg.com/markets/rates-bonds.

Figure 1. Policy Interest Rates in Eight Advanced Economies



Source: Bloomberg.

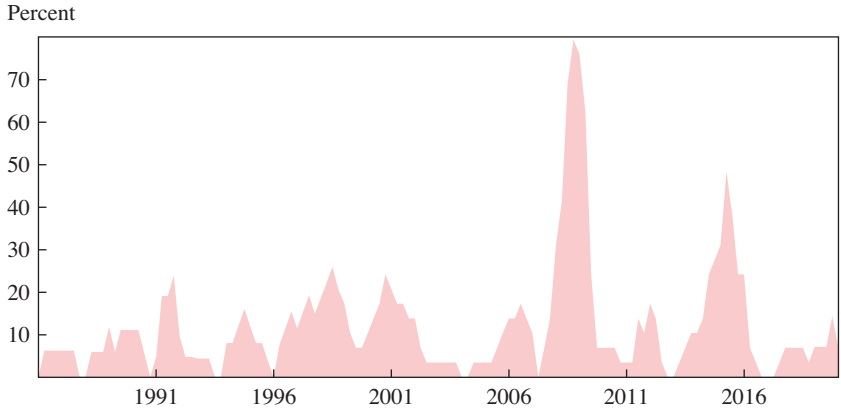
Note: US policy rate is the average of the range set by the Federal Reserve. Euro area is the interest rate on the European Central Bank's main refinancing operations. Data from January 1, 2021, through September 22, 2023.

was unexpected when countries were accumulating debt and making other financing decisions.[4] This also highlights the extent of surprise over a longer period than the short windows around monetary policy announcements that are often the focus of empirical analysis. Given that the surprise component of US monetary policy tends to have large spillover effects, this makes it even more noteworthy that the impact of this recent tightening in monetary policy on emerging markets has been muted.

This improved resilience of emerging markets, however, is not a new phenomenon and started well before the 2022 tightening in US monetary policy. For example, in 2020 when COVID-19 evolved into a global pandemic, risk spreads spiked, financial markets froze up, and emerging markets managed to avoid a series of financial crises and contagion (as was widely predicted by a number of economists). Granted, many emerging markets suffered sharp contractions in activity and major health challenges, as did most of the world, but many emerging markets were also much more resilient than expected. Let me provide two examples.

4. Federal Reserve Bank of New York, "Effective Federal Funds Rate," https://www.newyorkfed.org/markets/reference-rates/effr.

Figure 2. Share of Emerging Markets Experiencing a Sudden Stop



Source: Reproduced from Forbes and Warnock (2021) with permission from Elsevier.

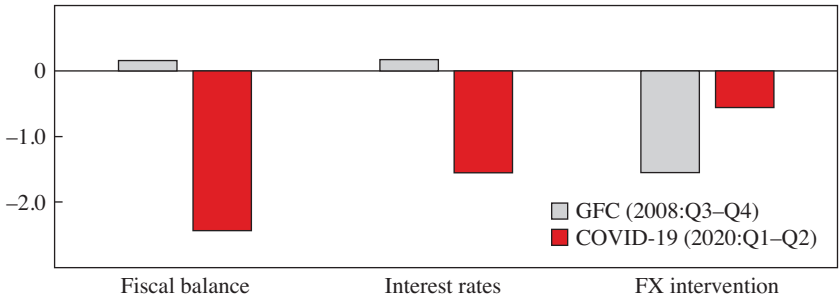
Note: Data from 1985:Q4 through 2020:Q2. A sudden stop is defined as a sharp decrease in gross capital inflows by foreigners relative to a country-specific historic average. See Forbes and Warnock (2012, 2021) for details on the methodology.

First, as the pandemic spread and global risk measures spiked (with some even higher than during the 2008 global financial crisis), emerging markets did not experience a wave of sudden stops in capital flows. Figure 2 replicates a graph from Forbes and Warnock (2021) showing the share of emerging markets experiencing a sudden stop in capital flows from foreigners from 1985 through the middle of 2020.[5] Only about 10 percent of emerging markets experienced a sudden stop in the first two quarters of 2020, well below the approximately 80 percent during the 2008 crisis. Forbes and Warnock (2021) describe the pattern of global capital flows after 2008 as more "ripples" than "waves." Their empirical analysis shows that capital flows became less sensitive to changes in global shocks (including risk measures, global growth, and US interest rates) after 2008.[6] They suggest that changes in the global financial system (e.g., tighter macroprudential regulation and reduced cross-border bank flows) have likely contributed to this improved resilience of emerging market capital flows for well over a decade.

5. More specifically, a sudden stop is defined as a sharp decrease in gross capital inflows by foreigners relative to a country-specific historic average. See Forbes and Warnock (2012, 2021) for details on the methodology.

6. Goldberg and Krogstrup (2018) and Avdjiev and others (2020) also find a reduced sensitivity of capital flows to global shocks in the 2010s relative to earlier periods.

Figure 3. Initial Policy Responses to COVID-19 in Emerging Markets



Source: Based on data compiled for Bergant and Forbes (2023).

Note: Initial policy response during the global financial crisis (GFC) is defined as 2008:Q3–Q4 and for COVID-19 as 2020:Q1–Q2. Fiscal balance is the change in the primary fiscal balance relative to GDP. Interest rates are the change in the policy interest rate. FX intervention is the change in FX reserves relative to per capita GDP. See Bergant and Forbes (2023) for a discussion of data sources and methodology.

A second, and related, example of this increased resilience from before the 2022 US monetary tightening is the greater ability of emerging markets to use countercyclical tools to support their economies in the face of global shocks—such as lowering interest rates and increasing government spending. This is in sharp contrast to many historical risk-off shocks (including increases in US interest rates) when emerging markets had to raise their domestic policy interest rate and reduce government spending in order to stabilize the exchange rate and capital flows. Figure 3 provides an example of this increased policy flexibility from Bergant and Forbes (2023).[7] During the initial phase of the COVID-19 pandemic (from 2020:Q1–Q2), emerging markets were able to lower interest rates and increase fiscal deficits to support their economies, a sharp contrast to what occurred during the initial phase of the 2008 global financial crisis (2008:Q3–Q4). Emerging markets also relied less on using reserves to support their exchange rates during the 2020 episode, a sharp contrast to the much larger reserve outflows during the 2008 crisis. In fact, by the end of 2020, even as COVID-19 still raged around the globe, many emerging markets began accumulating reserves as capital flows returned and began to worry about exchange rate appreciations that could damage competitiveness—a sharp reversal from the usual concerns about depreciations that traditionally occurred during risk-off shocks.

7. The fiscal response is measured as the change in the primary fiscal balance relative to GDP, and the monetary response is the change in the policy interest rate. The change in reserves is the change in FX reserves for exchange rate management relative to GDP per capita. Responses are for the first two quarters of each episode.

What explains this improved resilience of emerging markets since 2008—whether assessed by the reduced occurrence of sudden stops in capital flows, emerging markets' greater ability to use countercyclical policy tools, or their reduced vulnerability to the 2022–2023 US interest rate hikes? There are a number of possible explanations that could have played a role in at least a subset of major emerging markets.

- Reduced current account deficits—a vulnerability that received substantial attention during the taper tantrum as investors focused on the vulnerabilities related to large current account deficits in the Fragile Five.[8]
- Smaller aggregate volumes of gross capital inflows—especially of the more volatile types of flows, which could reduce vulnerability to risk shocks that affect global capital flows.
- Larger reserve stockpiles—which could be used to reduce exchange rate volatility (and the corresponding amplification effects through FX mismatches) as well as build investor confidence in the country's ability to manage shocks.
- More flexible exchange rates—which facilitate the adjustment to shocks and tend to increase the use of FX hedging so that entities can better withstand exchange rate movements.
- Stronger macroprudential regulations that have bolstered reserves and liquidity management in banks, making them more resilient to shocks and less likely to amplify shocks across the broader economy.[9]
- Improved credibility of monetary policy—which has allowed central banks to use monetary policy countercyclically to stabilize output and employment without increasing fears of price instability.
- Reduced exposure to FX debt, so that countries are less vulnerable to exchange rate movements.

I could make a strong case for why all of these changes (and others) have contributed to the greater resilience of many emerging markets to a range of shocks. Many of these changes are interrelated. This paper analyzes the last two potential explanations.

CONCERNS: DATA LIMITATIONS AND OMITTED VARIABLES In order to test whether improved monetary policy credibility and reduced FX borrowing have bolstered emerging market resilience, the authors focus on two data

8. The so-called Fragile Five were Brazil, India, Indonesia, Turkey, and South Africa.

9. See Forbes (2021) for evidence of the tightening in macroprudential regulations in emerging markets and how this has increased resilience to shocks.

sets. The first—on credibility—is a new measure constructed by Unsal, Papageorgiou, and Garbers (2022). It covers about fifty countries—which is very good country coverage—and the aggregate statistics summarizing key aspects of the data look logical. The data set is currently confidential, however, so it is difficult to get a good sense of the strengths and weaknesses of the data. Hopefully the authors will be able to share the data set at some point in the future as it looks promising and could be used for a range of applications.

The data used to analyze the second key variable of interest—FX exposure of the nonfinancial corporate sector from Kalemli-Özcan, Liu, and Shim (2021)—are also a logical start for capturing the vulnerability of one sector of a country to exchange rate movements and changes in risk premia. But I also would have liked to see the analysis repeated with other measures of FX exposure for several reasons. First, the current measure has very limited sample coverage, which presents challenges for the empirical methodology. Second, the current measure could miss important aspects of country vulnerability, as it does not include FX exposure of the nonbank financial sector (which has increased sharply and is a major focus of concern in the international financial institutions), exposure to FX other than US dollars, or the FX exposure of the banking sector.[10] Finally, it would also be useful to analyze FX exposure relative to the size of the economy and to its overall financial sector, rather than just as a share of outstanding credit, as countries with low debt levels may not be vulnerable even if a high share of these low debt levels are in FX. The bottom line—several alternate measures of FX exposure are available (albeit each has its advantages and disadvantages), and it should be straightforward to extend the analysis using other FX measures to see if the results are robust. This is particularly important as the FX results currently reported with the small sample do not appear to be robust across sensitivity tests (such as countries with more FX debt having better GDP performance in figure A5 in the online appendix).

This limited country coverage also raises a number of additional questions. How important are the outliers in driving the results—especially as two of the limited set of countries with FX data are Argentina and Turkey, countries that have had extremely volatile macroeconomic performance? And with such a limited sample, is there any way to control for other factors affecting

10. Although macroprudential regulations in most countries require banks to be hedged against FX exposure (as argued in the paper), banks can still be exposed through gross positions and through loans to entities that are not hedged, including nonbank financial institutions. See Forbes, Friedrich, and Reinhardt (2023) for evidence that banks were still vulnerable to FX exposure in 2020.

resilience to identify the individual contributions of the variables discussed above that also could be driving the increased resilience of emerging markets? For example, what is the role of tighter macroprudential regulations, changes to the nature of global capital flows, reduced current account deficits, more flexible exchange rates, and so on? In addition to these widely shared improvements, countries with more monetary policy credibility also likely have stronger institutional frameworks, higher income levels, stronger social safety nets, stronger macroprudential regulations, more stable inflation expectations, and more. Granted, many of these variables are endogenous (e.g., more credibility likely stabilized inflation expectations), but it is difficult to disentangle cause and effect in the current framework and with such a limited sample.

Put slightly differently, there is a strong negative correlation between monetary policy credibility and FX exposure (as the authors point out and as shown in an earlier draft). These variables are also correlated with other variables that could be important in bolstering resilience. For example, consider Chile—a country with strong monetary policy credibility and low share of FX debt. Chile has also been fairly resilient to higher US rates. But what explains this resilience? Is it the credibility of Banco Central de Chile? Or Chile's low share of FX debt? Or its low overall external borrowing and strong net foreign asset position (with its foreign asset positions often buffering shocks to foreign capital flows)? Or its strong institutions and rule of law? Or is it Chile's heavy exposure to copper and mining—for which the price has rebounded since 2021 around the same time the United States raised interest rates? Emerging market resilience after the COVID-19 pandemic could be driven by a number of factors—and sorting out the different influences will require an econometric approach that can better identify the different influences (and likely require some combination of a larger sample size and more time having passed to understand the 2021–2022 period).

FINAL THOUGHTS The question posed in this paper is important: why have emerging markets been fairly resilient (albeit with some prominent exceptions) as the United States raised interest rates much faster and to much higher levels than anyone expected even a year after the pandemic began? The authors focus on two potential explanations—improved monetary policy credibility and reduced FX exposure. I agree with their conclusions that both of these are key parts of the story. But there is also probably more to the story. Emerging market resilience has improved over a number of years and in response to a range of shocks. Using the 2022–2023 period of sharp increases in US interest rates to better understand which factors are

behind this resilience is worthwhile, but also challenging today due to the short time period combined with limited data for one of the key variables. This makes it impossible to control for omitted variables and to disentangle the many forces at play. I look forward to further iterations of this paper and more work on this topic to better understand these issues. The answer is critically important to provide guidance for how countries can best improve their resilience in the future—especially if we are entering an era of higher interest rates for an extended period.

REFERENCES FOR THE FORBES COMMENT

Avdjiev, Stefan, Leonardo Gambacorta, Linda S. Goldberg, and Stefano Schiaffi. 2020. "The Shifting Drivers of Global Liquidity." *Journal of International Economics* 125:103324.

Bergant, Katharina, and Kristin J. Forbes. 2023. "Policy Packages and Policy Space: Lessons from COVID-19." *European Economic Review* 158:104499.

Fan, Jingting, and Şebnem Kalemli-Özcan. 2016. "Emergence of Asia: Reforms, Corporate Savings, and Global Imbalances." *IMF Economic Review* 64, no. 2: 239–67.

Forbes, Kristin J. 2021. "The International Aspects of Macroprudential Policy." *Annual Review of Economics* 13:203–28.

Forbes, Kristin J., Christian Friedrich, and Dennis Reinhardt. 2023. "Stress Relief? Funding Structures and Resilience to the Covid Shock." *Journal of Monetary Economics* 137:47–81.

Forbes, Kristin J., and Francis E. Warnock. 2012. "Capital Flow Waves: Surges, Stops, Flight, and Retrenchment." *Journal of International Economics* 88, no. 2: 235–51.

Forbes, Kristin J., and Francis E. Warnock. 2021. "Capital Flow Waves—or Ripples? Extreme Capital Flow Movements since the Crisis." *Journal of International Money and Finance* 116:102394.

Goldberg, Linda S., and Signe Krogstrup. 2018. "International Capital Flow Pressures." Working Paper 24286. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w24286.

Kalemli-Özcan, Şebnem, Xiaoxi Liu, and Ilhyock Shim. 2021. "Exchange Rate Fluctuations and Firm Leverage." *IMF Economic Review* 69, no. 1: 90–121.

Unsal, Filiz D., Chris Papageorgiou, and Hendre Garbers. 2022. "Monetary Policy Frameworks: An Index and New Evidence." Working Paper 2022/022. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/01/28/Monetary-Policy-Frameworks-An-Index-and-New-Evidence-512228.

COMMENT BY

GIAN MARIA MILESI-FERRETTI This is a timely and interesting paper, which complements the excellent contribution to the Fall 2022 *BPEA* Conference by Obstfeld and Zhou (2022). Obstfeld and Zhou focused on

episodes of dollar appreciation and their impact on emerging market economies, while this paper by Kalemli-Özcan and Unsal focuses on the impact of US monetary policy tightening on the same set of countries. Rapidly rising US interest rates have in the past generated financial stress in the rest of the world, particularly in emerging and developing economies. The classic example remains the 1982 debt crisis, when the high US interest rates under the Volcker disinflation contributed to many external crises in emerging market economies, accompanied by severe GDP contractions. Many of these countries effectively regained access to global capital markets only in the early 1990s. Between February 1994 and February 1995, the Federal Reserve raised short-term interest rates by roughly 3 percentage points, and long-term interest rates went up by 1.5 percentage points.[1] The tightening led to a collapse of the Mexican peso—the country needed an international bailout to stave off default. The shock reverberated in Argentina as well, but this time there was no wider wave of emerging market crises. There also have been US monetary policy tightening episodes not associated with macroeconomic distress in emerging market economies—notably when the Federal Reserve raised interest rates from 1 percent to 5.25 percent between 2004 and 2006, as the United States and the global economy were staging a strong recovery.[2] The very rapid tightening of US monetary policy in 2022–2023 is an excellent moment to revisit the evidence.

One natural question is whether dollar appreciation and US monetary policy tightening are two faces of the same coin. In fact, they are correlated but not one and the same. The dollar appreciates during periods of rising global risk aversion, which can be periods of monetary policy easing (think of the global financial crisis). In contrast, there can be periods of substantial US monetary policy tightening (for instance 2004–2006) during which the dollar does not appreciate, as strong global demand and risk-taking reduce the importance of safe-haven factors.

Kalemli-Özcan and Unsal highlight two channels through which US monetary policy tightening can have repercussions in other countries. The first is the trade channel: to the extent that US monetary policy tightening is associated with currency depreciation vis-à-vis the US dollar, it could provide a boost to net exports. The authors argue that the existing evidence goes against the notion of an expansionary effect of exchange rate depreciations, in light of US dollar pricing and other factors. The second channel,

1. Board of Governors of the Federal Reserve System, "Selected Interest Rates (Daily)—H.15," https://www.federalreserve.gov/releases/h15/.

2. FRED, "Federal Funds Effective Rate," https://fred.stlouisfed.org/series/FEDFUNDS.

and the more salient one in discriminating between advanced economies and emerging markets, is the financial channel. Here the shedding of risky assets by global investors in response to tighter global financial conditions affects emerging markets more severely than advanced economies, as their credit ratings are generally lower and their risk profile higher.

But which factors are associated with the vulnerability to US monetary policy surprises? The authors focus on two key factors: monetary policy credibility and debt liabilities denominated in foreign currency. Their hypothesis is that rising monetary policy credibility and reduced foreign exchange exposures have increased the resilience of emerging market economies to spillovers from US monetary policy tightening. With regard to the challenging issue of measuring monetary policy credibility, a valuable innovation of the paper is the use of a very detailed index of monetary policy frameworks presented in Unsal, Papageorgiou, and Garbers (2022). This index is constructed by analyzing central banks' laws and websites for fifty advanced economies, emerging markets, and low-income developing countries, from 2007 to 2018, and focuses in particular on independence and accountability, policy and operational strategy, and communications. Once made public the data will be widely used in the profession.

The authors' findings for the period 1990–2019 are generally sensible. They underscore how emerging market economies are more severely affected by US monetary tightening than advanced economies; how, among emerging market economies, those with more credible monetary policy institutions are better able to cushion the impact of tightening global financial conditions on the domestic economy; and how US monetary policy tightening affects more severely those emerging market economies with balance sheet vulnerabilities in the form of high foreign exchange exposures.

Overall, the authors argue, emerging markets are in a better position now to deal with tighter global financial conditions than they were in previous decades, as they have strengthened their monetary policy institutions and policy frameworks and reduced their foreign exchange exposures. While the resilience to the post-COVID-19 US monetary policy tightening episode is consistent with this thesis, the shortness of the sample period complicates the task of distinguishing among different hypotheses.

I agree with the authors' general assessment, as I view the strengthening of monetary policy frameworks and the reduction of foreign exchange exposures as essential in explaining increased resilience to external shocks in emerging market economies. But there are other important aspects of emerging market policies and institutions that have contributed to increased resilience: more flexible exchange rates, stronger fiscal frameworks, improved

net external positions, and macroprudential regulation and supervision come to mind. On the investor side, with increased financial integration there was arguably an increase in investors' ability and willingness to differentiate across countries with different vulnerabilities. Given the correlation across many of these indicators, a formal pecking order is difficult to establish, but the variables the authors consider are certainly very important.

My discussion of the paper focuses primarily on three broad themes. The first is when did emerging markets become more resilient—I will argue that this has been an ongoing process within the first sample period the authors use (1990–2019), which was already bearing fruit well before the current episode. The second theme is the difficulty in drawing general inferences given the use of a changing mix of countries in the empirical analysis. The third is the strength of the empirical evidence presented on the role of foreign exchange exposures. I also discuss briefly the interpretation of the resilience to the latest monetary policy tightening episode and the measurement of such tightening in the empirical analysis.

WHEN DID EMERGING MARKETS BECOME MORE RESILIENT? The main focus of the paper is on the comparison between the response to monetary policy shocks in the period 1990–2019 and in the tightening episode occurring after the COVID-19 shock. The authors show how countries with different average characteristics in terms of central bank credibility and foreign exchange exposure responded to monetary policy shocks during the entire pre-COVID-19 period, with stable coefficients throughout (implying a similar response of all variables to US monetary policy shocks during these three decades).

However, the strengthening of emerging market balance sheets and monetary policy frameworks has been a gradual process that was already bearing fruit long before the COVID-19 shock. And indeed the paper highlights a process of rising resilience, starting in the 1990s: the monetary policy credibility index—which increases notably for emerging market economies between 2007 and 2021—corroborates this view. Emerging market crises have declined substantially in frequency since the early 2000s. During the global financial crisis, a few economies in Central and Eastern Europe (notably Hungary, Latvia, and Romania) had to rely on IMF programs, but elsewhere the incidence of crises was limited, especially when considering the depth of that global downturn. To be sure, external shocks—including at times US monetary policy tightening—had an impact on these economies, but such impact has been increasingly tempered by more resilient policy frameworks. The taper tantrum starting in May 2013—which the paper uses to illustrate the different responses to US monetary policy shocks between

Canada and Mexico—provides a good example of this. The shock generated sharp currency depreciations and large portfolio outflows from a number of emerging economies, including in particular a group called, at the time, the Fragile Five (Brazil, India, Indonesia, South Africa, and Turkey). However, the impact faded later in the year, and none of the affected countries experienced even a single quarter of negative growth in 2013—a big contrast with the deep recessions of the 1980s, the Tequila Crisis in Mexico, and the Asian crisis of 1997.

For these reasons it would be interesting to explore whether this process of increased resilience is supported by evidence on the response of emerging market economies to US monetary policy shocks between the earlier and the latter part of the sample.[3] This would also strengthen the case for the role of improvements in monetary policy frameworks within countries, since the evidence presented in the text relies on cross-sectional differences and the robustness check in the online appendix combines cross-sectional differences with time series evidence.

COUNTRY GROUPS The baseline results highlighting differences between advanced economies and emerging markets rely on a sample of fifty-nine countries. The breadth of the sample shows the thoroughness of the authors in establishing important stylized facts. At the same time, however, a number of countries in this specific sample have characteristics that differ to an important extent from those of the main emerging markets. Specifically, there are some countries with lower incomes or limited integration to global financial markets for a good part of the sample (for instance Albania, Armenia, Azerbaijan, and Belarus), current euro area members (Latvia, Malta, and the Slovak Republic), and hard pegs such as Bulgaria (after a high inflation period in the early 1990s). These countries are not part of the subsequent analysis, which explores differences in the reactions of emerging market economies to US monetary policy shocks depending on their monetary policy credibility and foreign exchange exposures.

While the authors have commendably undertaken a vast array of robustness exercises, I would have found it useful to establish the key stylized facts on the basis of a sample which is consistent across the paper, since data on the monetary policy credibility index are available for all the main emerging markets in terms of size and global importance. One important reason is that the assumption of a common coefficient across countries in the response of macroeconomic and financial variables to US interest rate

3. Ideally the sample would start a decade earlier, so as to encompass the debt crisis, but data challenges would be daunting.

shocks becomes harder to defend as heterogeneity as the level of GDP, financial development, and institutional frameworks increases. The US monetary policy tightening during the post-COVID-19 period provides a very useful illustration. While the largest and most developed emerging market economies fared well, a number of countries with weaker policy frameworks, such as Egypt, Pakistan, Sri Lanka, and Tunisia, have experienced severe market pressures or, in the case of Sri Lanka, a painful default.

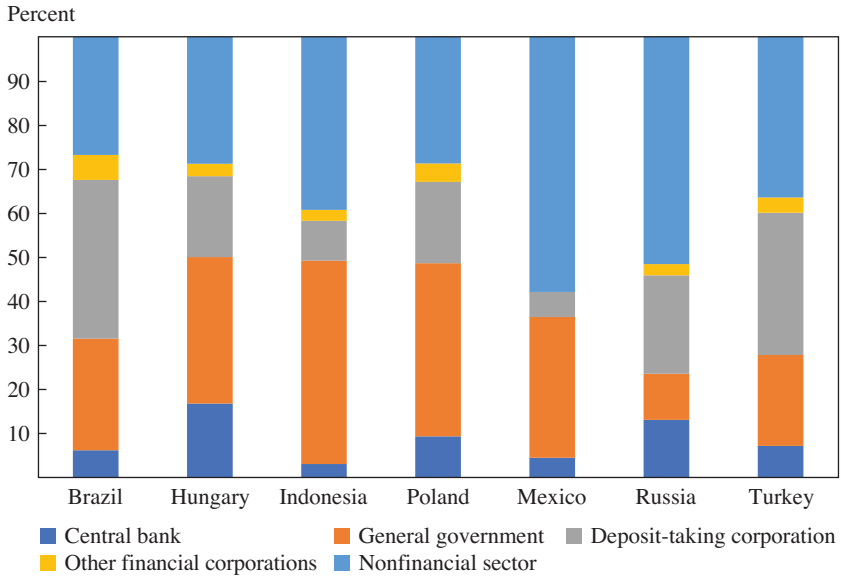
The most severe limitation in terms of data availability comes from the analysis of foreign exchange exposures. The Bank for International Settlements data used for this exercise on credit in foreign currency to the nonfinancial sector are available for only fifteen countries. These do not include countries in Central and Eastern Europe (with the exception of Russia) in which foreign exchange exposures were particularly important around the time of the global financial crisis—for instance in Hungary and Poland through mortgages denominated in currencies such as Swiss francs (Dizikes 2022; Minder 2022). The limited sample complicates the task of exploring differences in emerging market reactions to shocks depending on such exposures. Also, the strong negative correlation between the measure of foreign exchange exposure and the monetary policy credibility index (documented in the paper) raises questions as to whether the results for foreign exchange exposures could be capturing differences across countries due to the strength of monetary policy frameworks.

MEASURING FOREIGN EXCHANGE EXPOSURES Changes in foreign exchange exposures have been a crucial element in strengthening the resilience of emerging market economies. Since the early 1990s, their foreign exchange reserves have been rising, the composition of external liabilities in emerging market economies has shifted away from external debt toward foreign direct investment, and during the past two decades holdings by foreign investors of domestic currency government bonds have increased.[4] As a result, in the main emerging market economies, currency depreciations, while still costly, now improve the net external position, since the domestic economy is a net creditor in foreign exchange instruments.

Unfortunately there are many definitions of foreign exchange exposures (gross versus net, hedged versus unhedged, total versus vis-à-vis nonresidents) and no perfect comprehensive data set robustly based on microeconomic data. The variable chosen in the paper is a comprehensive measure of total foreign exchange exposure for the nonfinancial corporate sector, but that specific sectoral coverage and its reduced cross-country

4. See, for instance, Lane and Milesi-Ferretti (2007, 2018) and Arslanalp and Tsuda (2014).

Figure 1. External Debt in Foreign Currency (excluding Intercompany Lending): Sectoral Shares, 2021



Source: IMF Balance of Payments Statistics.

availability are important limitations. The authors argue that the currency exposure of banks is hedged. This may be the case in recent years for the largest emerging markets, but it is unlikely to be the case across the board for a protracted period of the thirty-year sample under consideration. But importantly, the definition also omits borrowing in foreign currency by the government, which is still important in most emerging market economies in the sample.

Figure 1 illustrates this point by making use of official data on the currency composition of external debt liabilities for the year 2021, published in the IMF Balance of Payments Statistics.[5] Furthermore, the relevance of borrowing by the nonfinancial corporate sector is higher for the more developed emerging markets (such as those shown here). For others, the role of government is even larger. What are the alternatives? The authors are reluctant to use data on the currency composition of external debt liabilities by Bénétrix and others (2019) because they do include FX liabilities by other

5. IMF, "Balance of Payments and International Investment Position Statistics (BOP/IIP)," https://data.imf.org/?sk=7a51304b-6426-40c0-83dd-ca473ca1fd52&sid=1542633711584.

sectors as well, and hence cannot single out the role of nonfinancial corporate sector liabilities in foreign currency. As argued above, I don't see this broader definition as a weakness. In the online appendix the authors run a robustness check using data on total external debt from the same source (as opposed to external debt denominated in foreign currency), but the results do not seem to show differences between high-debt and low-debt countries. A very recent paper by Allen, Gautam, and Juvenal (2023) updates and improves the Bénétrix and others (2019) currency composition data for countries' external balance sheets, supported by the publication of official data on such currency composition by a number of advanced economies and emerging markets (the data used in the construction of figure 1). It should provide a valuable tool for questions like the one used in this paper.

THE RESILIENCE TO POST-COVID-19 TIGHTENING OF US MONETARY POLICY The very limited time period limits the generality of the analysis of the post-COVID-19 period. The authors provide evidence that shows how during this episode the standard response to US monetary policy tightening (depreciation, rising risk premia, weaker GDP) has not materialized. This is undoubtedly correct for the main emerging markets. But with the data available so far, one cannot establish that this increased resilience is explained by stronger monetary policy frameworks or reduced foreign exchange exposures. In addition to other aspects of increased resilience mentioned earlier, the strength of commodity prices and an ongoing process of monetary policy tightening in emerging market economies starting well in advance of monetary policy tightening in the United States are likely to play a role as well.

One important feature of this episode has been the differentiation in markets between emerging market economies with varying levels of vulnerability.[6] The main emerging markets have so far emerged unscathed from the episode while a number of others (including Argentina—see figure 16 in the paper) and several low-income countries have faced market pressures and, in several cases, outright crises. In fairness to the authors, it may be difficult to capture this differentiation in their data given that several emerging market countries facing external challenges (for instance Egypt, Sri Lanka, and Tunisia) are not in the sample of countries with monetary policy credibility data.

HOW TO MEASURE MONETARY POLICY TIGHTENING The measures of monetary policy tightening used in the paper are standard in the literature and

6. Figure 16 in the paper illustrates the impact of widening Argentina's CDS spreads on the average for all emerging market economies.

well explained. Furthermore, the authors have undertaken a variety of robustness exercises using alternative measures of monetary policy shocks, which go beyond those presented in the final paper. I am still left with a question, particularly salient in an episode like the one we just observed. Namely, do monetary policy surprises (measured as changes in interest rates during a narrow time window around the monetary policy announcement) convey all relevant information on the extent of "surprise tightening"? The surprises during the latest US monetary policy tightening episode (shown by the authors in a previous draft) are generally small—yet changes in the Federal Open Market Committee (FOMC) "dot plot" from the first increase in rates in March 2022 to later that year and the most recent ones have been quite dramatic—as illustrated in Kristin Forbes's comment (Forbes 2024). While one could argue that these changes were not "surprises" but were driven by macroeconomic developments during the period, it is plausible that the Federal Reserve communication—for instance, speeches, testimonies, interviews with journalists, and so on—has played an important role in shaping market expectations about future rates, even outside FOMC meeting dates. A good historical example is the taper tantrum episode: it would not appear as a monetary policy surprise since it was a reaction to congressional testimony, and not to an FOMC meeting.

In conclusion, this paper will certainly stimulate much additional work. The authors have made a strong effort to be comprehensive and show a variety of results from different samples and specifications. But fully addressing all the issues the paper raises calls for more research in this area—and I very much look forward to that.

REFERENCES FOR THE MILESI-FERRETTI COMMENT

Allen, Cian, Deepali Gautam, and Luciana Juvenal. 2023. "Currencies of External Balance Sheets." Working Paper 2023/237. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2023/11/18/Currencies-of-External-Balance-Sheets-541610\.

Arslanalp, Serkan, and Takahiro Tsuda. 2014. "Tracking Global Demand for Emerging Market Sovereign Debt." Working Paper 2014/039. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2016/12/31/Tracking-Global-Demand-for-Emerging-Market-Sovereign-Debt-41399.

Bénétrix, Agustin, Deepali Gautam, Luciana Juvenal, and Martin Schmitz. 2019. "Cross-Border Currency Exposures." Working Paper 2019/299. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2019/12/27/Cross-Border-Currency-Exposures-48876.

Dizikes, Peter. 2022. "How the Debt Crisis of 2008–09 Fueled Populist Politics, *MIT News*, June 16. https://news.mit.edu/2022/hungary-debt-crisis-populist-0616.

Forbes, Kristin J. 2024. Comment on "Global Transmission of Fed Hikes: The Role of Policy Credibility and Balance Sheets," by Şebnem Kalemli-Özcan and Filiz Unsal. In the present volume of *Brookings Papers on Economic Activity*.

Lane, Philip R., and Gian Maria Milesi-Ferretti. 2007. "The External Wealth of Nations Mark II: Revised and Extended Estimates of Foreign Assets and Liabilities, 1970–2004." *Journal of International Economics* 73, no. 2: 223–50.

Lane, Philip R., and Gian Maria Milesi-Ferretti. 2018. "The External Wealth of Nations Revisited: International Financial Integration in the Aftermath of the Global Financial Crisis." *IMF Economic Review* 66:189–222.

Minder, Raphael. 2022. "The Mortgage Time Bomb Ticking beneath Poland's Banks." *Financial Times*, November 12. https://www.ft.com/content/19235a7a-36f1-41a4-b58f-6adfdea53220.

Obstfeld, Maurice, and Haonan Zhou. 2022. "The Global Dollar Cycle." *Brookings Papers on Economic Activity*, Fall, 361–427.

Unsal, Filiz D., Chris Papageorgiou, and Hendre Garbers. 2022. "Monetary Policy Frameworks: An Index and New Evidence." Working Paper 2022/022. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/01/28/Monetary-Policy-Frameworks-An-Index-and-New-Evidence-512228.

GENERAL DISCUSSION Jonathan Pingle began the discussion by highlighting that in addition to the countries the authors consider in their paper, the US economy also remained unusually resilient to the 2022–2023 US Federal Reserve interest rate hikes. He asked to what extent this plays a role, noting that it would affect risk premia, risk sentiments, equity evaluations, business investment, and foreign direct investment. Pingle wondered whether there could have been additional factors affecting the resilience of emerging markets.

Jason Furman pointed out that one hypothesis explaining the lack of spillover posits that, due to the common shock element, emerging markets began to raise rates before the Federal Reserve decided to do so and likely would have even absent rate hikes in the United States. This stands in contrast to previous tightening cycles where emerging markets were less macroeconomically synchronized with the United States, leading to undesirable rate hikes in those economies. Furman inquired about the authors' thoughts on this and suggested controlling for a pooled common shock or using time fixed effects to address this possibility.

Donald Kohn similarly noted that many of the central banks in emerging markets raised rates before the Federal Reserve recognized the problem because many emerging markets now have the independence to do so. This independence has developed over time and ultimately protected their credibility and better insulated their economies.

Ayşegül Şahin brought up the transmission mechanism of Fed hikes in the authors' analysis. She asked whether the transmission of interest rate hikes was primarily through the trade channel and commodity markets or the financial channel, and whether the authors had a sense of the relative magnitudes of the transmission mechanisms.

In response to the observations about US resilience to Fed hikes, Şebnem Kalemli-Özcan commented that, rather than the response of the US economy, what is important is how global financial conditions responded to Fed hikes. She highlighted that the primary channel Fed hikes pass through is the risk sentiment of financial investors and how tight global financial conditions are. She emphasized the focus of the authors on these two factors for a given change in US interest rates. She also remarked that while not all changes in US interest rates affect risk sentiments, changes in risk sentiments can have a large impact on real macroeconomic variables.

Kalemli-Özcan further noted that their paper focuses on the financial channel because the trade channel, which they define as expenditure switching, works in a smoothing way, thus the effects of Fed hikes are not immediately realized in the trade channel. She also argued that the adverse effects of Fed hikes often materialize in the financial channel, rather than the trade channel, through changes in risk premia. Kalemli-Özcan pointed out that their model does control for changes in the trade channel, including the current account and capital account balances, among other trade-related variables.

Jordi Galí commented that the Gertler-Karadi shocks used in the paper may not constitute a pure exogenous shock but may have an endogenous component if the central banks have private information about the prospects of the economy that the financial markets do not.[1] He mentioned that if the relative importance of the endogenous and the exogenous components had changed over time, possibly due to improved, more systematic Federal Reserve policy, perhaps that could partly explain the authors' results.

1. Mark Gertler and Peter Karadi, "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics* 7, no. 1 (2015): 44–76.

He noted that what one may have interpreted as exogenous in the most recent tightening of monetary policy in the United States may instead have reflected the prospect of an improvement in the US economy—which may have also had positive impacts on emerging market economies through trade links, for example.

Kalemli-Özcan remarked that they experimented with different types of shocks in their analysis, including Bauer-Swanson, Nakamura-Steinsson, and Gertler-Karadi shocks, as well as additional risk sentiment measures.[2] In response to Galí's comment, she agreed that the sensitivities can be important domestically but noted that internationally the effects of a rate hike work similarly as long as the shock detects changes in the risk sentiment. She emphasized that it is not about the size of the monetary policy shock—not every shock will change the risk sentiment—but rather the extent to which the risk sentiment changes, arguing that the trade channel and other forms of linkages are not as important.

Caroline Hoxby inquired about the rise in external financing through foreign direct investment and other equity instruments and asked about the authors' thoughts on whether these could have played a role in emerging markets' resilience.

Steven Kamin agreed with the paper's findings but suggested the authors include the global financial market's reactions to the Federal Reserve tightening, noting that it spills into emerging markets by affecting global risk sentiment, thereby causing capital outflows from emerging markets. One way this can be observed is through US high-yield corporate spreads, which are highly correlated with emerging market dollar-bond credit spreads. Kamin highlighted that if the US financial conditions were to deteriorate in the coming years, it could lead to greater deterioration for emerging markets as well.

In response, Kalemli-Özcan affirmed that US high-yield corporate spreads did not drastically change during the recent Federal Reserve interest rate hikes and emphasized that this is corroborating evidence that the rapid tightening didn't create a risk-off shock.[3]

2. Michael D. Bauer and Eric T. Swanson, "A Reassessment of Monetary Policy Surprises and High-Frequency Identification," *NBER Macroeconomics Annual* 37, no. 1 (2023): 87–155; Emi Nakamura and Jón Steinsson, "High-Frequency Identification of Monetary Non-neutrality: The Information Effect," *Quarterly Journal of Economics* 133, no. 3 (2018): 1283–330.

3. Collin Martin, "High-Yield Bonds: Yields Are Up, but Risks Remain," Charles Schwab, August 31, 2023, https://www.schwab.com/learn/story/high-yield-bonds-yields-are-up-but-risks-remain.

Filiz Unsal discussed the confidentiality of the data. To measure the monetary policy credibility of countries, Kalemli-Özcan and Unsal collected data from central bank laws, websites, and communications. Using these data, they rated the credibility of each country using a framework developed by Unsal and colleagues.[4] She explained that this measure of monetary policy credibility extends beyond countries reaching their inflation targets, encompassing many aspects of the monetary policy-making process. Kalemli-Özcan further commented that the measure of monetary policy credibility can be summarized as "Do what you say, say what you do." This includes committing to price stability and being forthcoming about the methods of attaining goals. If a country were to make these commitments but then attempt to influence exchange rates and capital flows with the interest rate, it would not be considered credible monetary policy under their framework. Improvements in monetary policy credibility across emerging markets can be attributed in part to the use of macroprudential policies to manage debt denominated in foreign currency, resulting in decreased foreign debt.

Kalemli-Özcan also addressed questions posed by Kristin Forbes and Gian Maria Milesi-Ferretti during their discussant remarks. Milesi-Ferretti pointed out that some of the countries included in the initial sample had different reactions to the recent Fed hikes and different monetary policy regimes, in particular lower-income countries as well as Argentina and Saudi Arabia. Kalemli-Özcan explained that initially the authors included a sample of emerging markets and developing countries in line with Obstfeld and Zhou.[5] She noted that the response is late and heterogeneous for low-income countries and also agreed that the resiliency of countries to recent Fed hikes only applies to emerging markets. Kalemli-Özcan affirmed that they ended up dropping these countries from the sample.

In her presentation, Forbes discussed the paper's exclusion of nonbank financial sector foreign debt, which has gone up considerably in recent years due to tighter macroprudential policies, and the paper's focus on the nonfinancial private sector in the foreign exchange (FX) exposure data. She noted that this measure was restrictive and didn't have enough observations. Kalemli-Özcan responded that the vulnerability they attempted to

4. Filiz D. Unsal, Chris Papageorgiou, and Hendre Garbers, "Monetary Policy Frameworks: An Index and New Evidence," working paper 2022/022 (Washington: International Monetary Fund, 2022). https://www.imf.org/en/Publications/WP/Issues/2022/01/28/Monetary-Policy-Frameworks-An-Index-and-New-Evidence-512228.

5. Maurice Obstfeld and Haonan Zhou, "The Global Dollar Cycle," *Brookings Papers on Economic Activity*, Fall 2022, 361–427.

measure is the unhedged dollar debt in the private sector. She outlined that historically, during the Fed hikes, high levels of debt in the nonfinancial private sector of emerging markets led to economic contractions. Thus, countries would ideally seek to counteract the contraction by lowering interest rates, but at the same time countries needed to raise interest rates in line with the Federal Reserve to keep their currencies afloat. This is the vulnerability that the authors were attempting to capture.

In response to the discussant remarks about the FX exposure data, Kalemli-Özcan noted that they interacted the continuous FX exposure data in addition to high and low exposure categorical variable with the monetary policy shocks. Therefore, there is both a continuous and discrete aspect of the FX exposure variable. With the interacted regressor, the authors included a time fixed effect in the model, which controlled for commodity prices, oil prices, VIX, and other global financial variables. In terms of the time periods used in their paper, Kalemli-Özcan noted that capital outflows were much greater during the global financial crisis than in recent periods and affirmed that they could show the differences between periods.

RYAN A. DECKER
Federal Reserve Board

JOHN HALTIWANGER
University of Maryland

Surging Business Formation in the Pandemic: Causes and Consequences?

ABSTRACT Applications for new businesses surprisingly surged during the COVID-19 pandemic, rising the most in industries rooted in pandemic-era changes to work, lifestyle, and business. The unexpected surge in applications raised questions about whether a surge in actual new employer businesses would follow. Evidence now shows increased employer business entry with notable associated job creation; and industries and locations with the largest increase in applications have had accompanying large increases in employer business entry. We also observe a tight connection between the surge in applications and quits—or close proxies for quits—both at the national and the local level. Within major cities, applications, net establishment entry, and our quits proxy each exhibit a "donut pattern," with less growth in city centers than in the surrounding areas, and these patterns are closely related to patterns of work-from-home activity. Reallocation of jobs across firm age, firm size, industry, and geography groupings increased significantly. Relatedly, there is evidence of a pause of the pre-pandemic trend toward greater economic activity being concentrated at large and mature firms, but this development is quite modest in magnitude.

Conflict of Interest Disclosure: The authors did not receive financial support from any firm or person for this paper or from any firm or person with a financial or political interest in this paper. The authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper.

Brookings Papers on Economic Activity, Fall 2023: 249–302

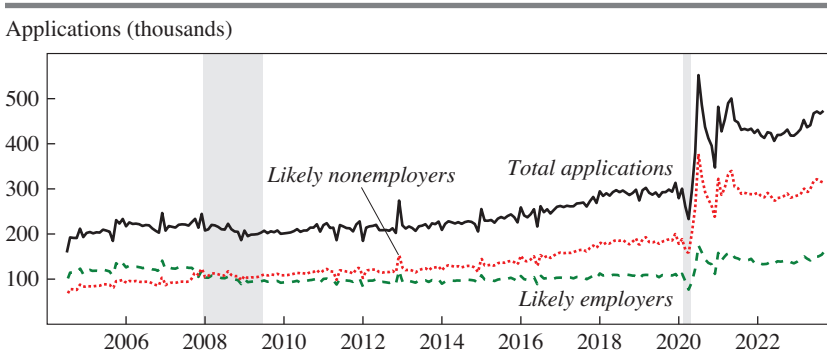
The US economic experience during the COVID-19 pandemic featured a surprising surge in applications for new businesses, shown in figure 1. After dropping in March and April of 2020, applications rose sharply, reaching an all-time high in July 2020; the series declined through the rest of 2020, then surged again in 2021, and have remained historically elevated through September 2023. These data received widespread attention amid high unemployment and broader economic volatility, in part because the surge was apparent even among "likely employers," that is, applications with characteristics that predict the hiring of workers and growth.[1] Monthly applications for likely employer businesses in September 2023 were more than 30 percent higher than the 2019 pace. Historically, there has been a tight relationship between likely employer business applications and true employer business formation, but questions have remained about whether the pandemic's surging applications would translate into actual employer businesses with broader macroeconomic implications.

In this paper, we describe noteworthy aspects of the surprising surge in applications that point to its genuine economic content. We then draw on a range of data sources to show that the surge in applications was followed—after some lag—by a surge in employer business creation: quarterly data on establishment entry rose substantially starting in the second quarter of 2021, while annual data on firm entry jumped in the year ending March 2022 (figure 2). Moreover, we document a close empirical relationship between applications and employer business entry across industry and geography, with hallmark patterns from the application data appearing in employer entry data. We relate the surge in business formation to pandemic labor market stories such as the Great Resignation, that is, the rise in worker quit rates starting in early 2021 (Rosenberg 2022). Finally, we describe the striking resilience of small and young firms through the pandemic period, and we highlight modest hints of a reversal of pre-pandemic trends in business dynamism—though we note that it is too early to declare an end to those trends.

This set of facts lends itself to a compelling narrative of pandemic business and labor market dynamics. The pandemic sparked rapid, dramatic changes to the composition of consumer demand and to preferences for work, lifestyle, and business; and these patterns continued to evolve into 2023. From the standpoint of potential entrepreneurs, these dramatic

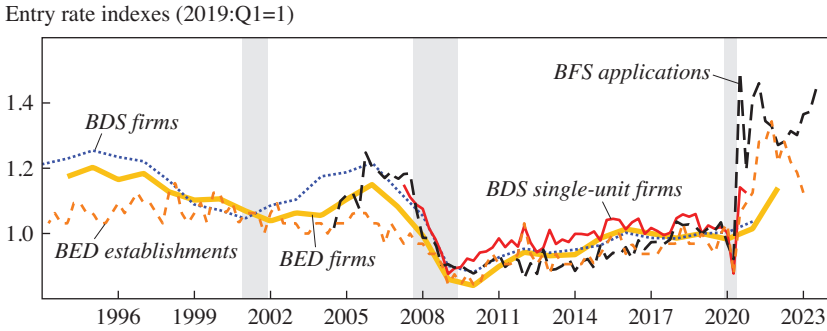
1. We more completely describe "likely employer" applications and the data from which they are derived in section I and online appendix A.

Figure 1. New Business Applications



Source: Census Bureau Business Formation Statistics.
Note: Seasonally adjusted. Total applications = BA series; likely employers = HBA (high-propensity applications) series; likely nonemployers is residual. Shaded areas indicate NBER recession dates.

Figure 2. New Business Entry and New Business Applications



Source: Business Dynamics Statistics (BDS); Business Employment Dynamics (BED); and Business Formation Statistics (BFS).
Note: BDS and BED annual firm births are age zero firms as of March. BFS applications are likely employers (the HBA series). All series expressed as rates except BFS. Quarterly series are seasonally adjusted. Gray bars indicate NBER recession dates (2001:Q1–2001:Q4, 2007:Q4–2009:Q2, 2020:Q1–2020:Q2).

changes presented opportunities—both to meet newly formed consumer and business needs and to change the career trajectories of the entrepreneurs themselves. Entrepreneurs made plans and applied to start businesses both early on and through the fall of 2023; some of these plans have resulted in new firms and establishments that hired workers in large numbers. Entrepreneurial opportunities and the demand for employees at these new firms appear to have played an important role in the Great

Resignation, as some quitting workers likely flowed toward new businesses (as either entrepreneurs or new hires). Taken together, these patterns imply significant economic restructuring across industry, geography, and the firm size and age distribution. The extent to which these changes will be long-lasting has yet to be seen.

The surge in applications started in the second half of 2020, but it has taken time to determine the implications for new employer (and nonemployer) businesses. One reason for the delay may be that the initial surge in the summer of 2020 was relatively short-lived, with the more sustained surge in applications commencing later—in early 2021. Moreover, likely employer applications take up to eight quarters to yield the first hire—even conditional on making that transition. And in the United States, data on the creation of actual employer businesses—that is, businesses with paid workers—are published with a lag since such measures derive from administrative data with long processing time. The timeliest data on new employer businesses are for *establishment* births from the Bureau of Labor Statistics (BLS) Business Employment Dynamics (BED); as of September 2023, BED data on establishment births are available through 2023:Q1, while BED data on (annual) *firm* births are available through March 2022.[2] The gold standard annual firm birth data from the Census Bureau Business Dynamic Statistics (BDS) are available through March 2021 for all firms, while quarterly data on single-establishment firms go through 2020:Q4. Between these and other sources, we now have sufficient data to characterize patterns of employer business formation and related job and worker flows in the pandemic.

We observe strong sectoral and geographic correlations between business applications and employer business entry (we measure the latter by either firms or establishments, and in either gross or net terms, depending on data availability). The rise in applications and employer entry is highly concentrated in a few industries that are conducive to pandemic patterns of work, lifestyle, and business (such as online retail and other high-tech industries), consistent with the changing sectoral structure of the economy. We also observe substantial spatial variation in the surge in applications and business entry, consistent with geographic restructuring. The surge in applications and business entry is especially notable in the South, with states such as Georgia standing out. Within large cities we observe a "donut

2. An establishment is a single business operating location—such as one's local Starbucks location—while a firm is a group of one or more establishments under a common tax identifier (in BLS measures) or under common operational control or ownership (in Census Bureau measures).

effect" with applications surging more in the suburbs of metropolitan areas than in central business districts.

The pandemic and its aftermath have been associated with increased churn of workers as found in (initially) elevated layoffs, many of them temporary (Cajner and others 2020), and, through much of the pandemic, elevated quits. We find a tight spatial correlation—at the state and county level—between surging business applications and quits (or excess separations, a close proxy for quits), with a much weaker correlation between applications and layoffs (or job destruction, a close proxy for layoffs). Among other possible explanations, these results are consistent with workers quitting their jobs to start or join new businesses—and much less consistent with job loss being a key driver of business formation.

This pandemic surge in entry occurred after decades of declining business dynamism in the United States. The pace of job reallocation had fallen by about 25 percent from the 1990s to just before the pandemic.[3] This decline in the pace of job reallocation was driven in part by the decline in employer business entry over this same time period, which can be seen in figure 2 or, for a longer view, online appendix figure E1; closely related is the shift of the firm distribution toward large and mature firms. While the sources of this decline have been widely debated in the literature, there is evidence that it has been associated with a decline in productivity-enhancing reallocation and is likely one of the factors underlying sluggish productivity growth in the United States since the early 2000s.[4]

3. US Census Bureau, "BDS Data: 2021 Business Dynamics Statistics Data Tables," https://www.census.gov/programs-surveys/bds/data.html. From the Business Dynamics Statistics (BDS) data, the average pace of job reallocation (job creation plus job destruction) in 1997–1999 was about 32 percent and in the 2017–2019 period about 24 percent.

4. As discussed in Davis and Haltiwanger (2014), there are likely both benign and adverse factors underlying this decline in business dynamism. However, as discussed in Decker, Haltiwanger, and others (2020), there has been a decline in the responsiveness of businesses to idiosyncratic productivity shocks and a widening of revenue productivity dispersion—both consistent with rising distortions and frictions in the economy. Alon and others (2018) present related evidence that the shift in activity to more mature firms has contributed to the decline in productivity growth. Moreover, Akcigit and Kerr (2018) and Acemoglu and others (2018) show evidence that young and small firms are more likely to make radical innovations, while mature incumbents make more incremental innovations in order to avoid cannibalizing their market share. Akcigit and Goldschlag (2023) present evidence that in the post-2000 period inventors are more likely to join large incumbents than young firms; moreover, they find that inventors who join large firms obtain higher earnings but are less innovative. They argue that this is due to strategic considerations for the same argument made above—to avoid cannibalizing their market share. De Loecker, Eeckhout, and Unger (2020) and Autor and others (2020) present evidence of rising markups associated with the shift to larger firms.

We show that the pandemic featured a surge in job reallocation, including reallocation between cells defined by industry, geography, firm size, and—especially—firm age. We also document a pandemic pause—and modest reversal—of the longer-run shift in activity toward large, mature businesses. The share of activity accounted for by young and small firms has ticked up; young and small firms exhibit a higher pace of dynamism than large and mature firms, so one might anticipate an ongoing increase in the pace of dynamism. In other words, we find early hints of a revival of business dynamism; but in many respects it is too early to ascertain whether a durable reversal of pre-pandemic trends is occurring. Such a reversal—that is, a persistent rise in the pace of reallocation and a substantial shift of activity away from large, mature firms—will require a long-lasting continuation of elevated business entry as well as substantial growth among at least a subset of the pandemic entrants.

It is useful to state our view of our contribution—and the limits to that contribution. A key contribution of our work is that we draw on a wide range of data sources: the Bureau of Labor Statistics' BED, Quarterly Census of Employment and Wages (QCEW), and Job Openings and Labor Turnover Survey (JOLTS), and the Census Bureau's Business Formation Statistics (BFS), BDS, and Quarterly Workforce Indicators (QWI). While none of these data sources alone can tell a comprehensive story of pandemic business entry, each contributes a different perspective in terms of timeliness, industry and geography detail, or measurement concept. We provide an initial assessment of the potential causes and consequences of the surge in business applications by supplying a rich set of empirical facts pointing to substantive pandemic economic stories, but we do not provide identified causal empirical results or new formal theory; rather, we hope our results can direct and discipline future causal analysis. We also hope our approach of exploiting an eclectic combination of data sets can help other researchers better understand the range of available business dynamics and labor market data that can inform timely analysis.

A study of actual application-to-employer transitions, post-entry dynamics, and job-to-job flows of workers must wait for the availability of administrative micro data.[5] Such micro data can also facilitate rigorous

5. Dinlersoz and others (2023) feature pre-pandemic cross-sectional analysis of the BFS micro data; it will be feasible to extend that work to the pandemic era once the administrative micro data tracking transitions and post-entry growth become available. This will require the confidential Longitudinal Business Database (LBD), which is currently available through March 2021.

causal analysis and provide additional empirical moments of relevance to theoretical investigations. Separately, while we focus on new employer businesses, the likely surge in new nonemployer businesses appears important and interesting as well; unfortunately, the nonemployer economy is measured with less detail and timeliness than the employer economy, so we leave that investigation for future work (but we provide some additional discussion near the end of this paper and in online appendix A).

Our work complements that of Fazio and others (2021), which documents similar aggregate patterns using zip code–level data on business registrations in eight states from the Startup Cartography Project; Fazio and others (2021) report striking time series relationships between pandemic fiscal stimulus and the registration surge and find that the surge was concentrated in zip codes with relatively high African American population and above-median income. They also find that the surge is apparent outside city centers within large cities; we show that this within-city pattern is apparent in county-level applications data for the United States as a whole, and we build on their earlier work by studying outcomes for net establishment entry and excess worker flows as well. Duguid and others (2023) document similar within-city patterns for retail establishments using credit card merchant data and relate these patterns to population flows and remote work considerations. We also expand on Decker and Haltiwanger (2022), in which we provided a first look at the relationships between business applications and establishment births (and exits) in official data and initially documented the increase in small firms' share of activity during the pandemic.[6]

In section I we briefly describe our main data sources, with much more detail in online appendix A. We review and document patterns of business applications in section II, then explore employer establishment and firm entry and their empirical relationship with applications in section III. We examine the relationship between worker churning—especially quits—and applications in section IV. In section V we document changes in the firm size and age distribution and consider implications for business dynamism. We take stock in section VI, then speculate about potential implications for the future in section VII.

6. An even earlier first look at the BFS surge in new business applications is in Haltiwanger (2022). This analysis focused on the surge in new business applications in the first year of the pandemic before data on actual employer business entry were available.

I. Data

We exploit a variety of data sources, all of which are publicly available tabulations. Online appendix A describes each source in detail; here we simply list our main sources with brief descriptions.

Business Formation Statistics (BFS), US Census Bureau: monthly data on IRS employer identification number (EIN) applications. All employer businesses and nonemployer corporations and partnerships must have an EIN, and many nonemployer sole proprietors choose to obtain one for business reasons. The total applications series (called "BA" in the BFS files) counts all EIN applications that are potential employer or nonemployer (zero-employee) businesses (this implies excluding applications for trusts, estates, and financial instruments). Our main interest is employer businesses; therefore, where possible we focus on what we call likely employer applications (high-propensity applications or "HBA" in the BFS files). This subset of the total applications series is based on Census Bureau modeling using application characteristics that have a high propensity for transitioning into an actual employer business with paid workers; these characteristics include planned hiring and corporate legal form, among others. However, at narrow levels of industry (three-digit NAICS) or geography (county) detail, only total applications are publicly available, so we use the total applications series as a proxy for our preferred likely employer series. As shown in figure 1 (and below at more disaggregated levels by industry and geography), total and likely employer applications have tracked each other closely in the pandemic, which mitigates concerns about using the total series as a proxy for likely employers where necessary. We use BFS series through September 2023.

The BFS also includes series that report, in any given time period, the number of applications that actually transition to genuinely new employer firms within four or eight quarters. These series use micro data linkages tying applications to actual employer firm births; the four-quarter and eight-quarter series are currently populated through 2019:Q4 and 2018:Q4, respectively, and relate to new employer firm micro data available through 2020:Q4. Since these transition series end relatively early (constrained by actual employer firm data timing in Census data), the BFS also features series for projected transitions at four- and eight-quarter horizons, where projections are based on application characteristics and include all applications (not just those labeled as likely employers). The motivation for the four- and eight-quarter horizons for actual and predicted transitions is that, as discussed further below, there is often a lag between applications

and transitions. An advantage of the projected series is that they take into account the full range of application characteristics (e.g., reason for application and detailed industry).[7]

Quarterly Census of Employment and Wages (QCEW), Bureau of Labor Statistics: quarterly establishment and employment counts by detailed industry and geography. The QCEW is derived from the main business register of the BLS and is based on state unemployment insurance administrative data. We use the QCEW to measure net establishment growth at the national, industry, and local (county) level. The QCEW micro data also underly the Business Employment Dynamics.

Business Employment Dynamics (BED), Bureau of Labor Statistics: quarterly data on establishment openings, closings, births, exits, expansions, and contractions, with associated job flows. The BED also features a research product with annual employment, firm, and establishment counts by firm age, where a firm is defined by an EIN. We use quarterly BED data extending through 2023:Q1 and annual firm age data through 2022:Q1. Importantly, in the BED, an establishment (firm) birth represents an establishment (firm) that did not previously exist; a new firm requires a new business application, while a new establishment of an existing firm does not require but may obtain a new EIN. Notably, new EINs acquired by existing firms would not count as employer firm transitions in the BFS four-quarter and eight-quarter transition series mentioned above but may appear as new establishments (or firms) in BED data.

Quarterly Workforce Indicators (QWI), US Census Bureau: quarterly data on employment and job and worker flows (i.e., hires and separations) by firm age with detailed industry (four-digit NAICS) and geography (county) tabulations. The QWI is the public-use version of the Longitudinal Employer-Household Dynamics (LEHD) data based on state unemployment insurance records and collected on a state-by-state basis; we use a balanced panel of forty-five states that covers just over 80 percent of private employment as of 2020. These data extend through 2022:Q2.

7. The likely employer series uses a more limited set of characteristics without the characteristic-specific loading factors from the estimated projection model that underlies the projected series. The projected series are by design a more reliable predictor of actual employer business formation, especially at the sector level. We include additional discussion of this issue in online appendix B. We primarily use the likely employer series in the main text since it is more transparent and because it is more comparable to the total applications series we must use for analysis of detailed industry or geography patterns, and there is generally a tight relationship between likely employer and the projected business formation series.

Job Openings and Labor Turnover Survey (JOLTS), Bureau of Labor Statistics: monthly survey-based estimates of hires, separations, quits, and layoffs with state-level detail. We use JOLTS data through September 2023 with a focus on quits and layoffs.

American Community Survey (ACS), US Census Bureau: annual survey-based data on work-from-home (WFH) prevalence for large counties. ACS data are available in two samples: five-year samples including the entire United States, and one-year samples including large counties. We use the one-year sample for 2019–2021 and focus on changes in WFH prevalence across counties within large cities. ACS WFH measures are based on location of worker residence; we discuss existing literature on WFH using other data (Hansen and others 2023) in online appendix A.

Additionally, we use data from the Census Bureau's Business Dynamics Statistics (BDS) in certain figures (e.g., figure 2); these data do not currently cover the pandemic period, so we do not use them in most of the exercises that follow. In online appendix A, we provide a discussion of the BDS and its relation to the BLS data sources listed above.

II. Business Application Patterns

II.A. The Early Pandemic Period

At the onset of the pandemic, plummeting weekly business application and registration data received widespread attention (Fazio, Guzman, and Stern 2020; Haltiwanger 2020; Federal Reserve System Board of Governors 2020).[8] But, as shown in figure 1, applications quickly recovered and surged to historic levels in July 2020. The surge is apparent in every application series, including total applications and likely employer applications (both shown in figure 1) as well as applications with planned wages and applications for corporations.[9] Applications did fall off in August 2020 through December 2020 (albeit still higher in December 2020 than prior to the pandemic) but then surged again in early 2021. This second wave has been more resilient, with monthly likely employer applications in 2023 so

8. See also Fairlie (2020), who tracks the number of business owners in Current Population Survey (CPS) data. Cognizant of challenges associated with measuring self-employment in CPS data (Abraham and others 2021), we do not explore CPS self-employment data in this paper.

9. Fazio and others (2021) similarly find surging business registrations for LLCs, partnerships, and corporations; interestingly, they find no surge among Delaware corporate forms preferred by venture capitalists.

far averaging about 30 percent higher than the 2019 pace. Total applications are about 40 percent higher in 2023 relative to 2019, reflecting the even larger surge of likely nonemployers.

The sharp rise in the likely employers series is in stark contrast to the previous recession. Dinlersoz and others (2021) and Haltiwanger (2022) explore this comparison in detail; here we note that the decline in total applications seen in the Great Recession was driven by the likely employer series, while the likely nonemployer series was roughly flat in that episode.[10] Flat or even rising nonemployer entrepreneurship during a recession can easily be rationalized in light of lack of opportunities for wage and salary employment, which may push many individuals into self-employment activities out of necessity; and, indeed, one plausible explanation for the pandemic surge in applications was that unemployment was elevated in the wake of spring 2020 shutdowns. But rising employer entrepreneurship is more difficult to understand, as businesses hiring employees are more likely to be pursuing genuine entrepreneurial opportunities; hence, the stark difference in likely employer behavior between the pandemic recession and the prior recession is all the more striking. And the pandemic surge in applications has persisted even as unemployment has fallen toward historic lows.

A number of factors could help account for the surge in applications for likely employers in the pandemic compared to the drop of likely employer applications and employer start-ups in the Great Recession. The pandemic provided new market opportunities given the changing nature of consumer demand and of work and lifestyles, and financial conditions—including house prices—were robust compared to the Great Recession (at least through early 2022). The potentially supportive role of stimulus programs—which included sizable support for aggregate demand and household balance sheets—is an open question. The US federal and state governments implemented a wide range of fiscal support programs which could have had myriad effects on business formation; one example is the expansion of unemployment insurance benefits, which Choi and others (2023) find had a positive effect on business applications. On the other hand, programs like the Paycheck Protection Program (PPP)—along with other business

10. Data on actual nonemployer activity during the Great Recession broadly confirm the relative resilience of the likely nonemployer applications data in that episode. The total number of actual nonemployer businesses declined just 1.6 percent between 2007 and 2008 but fully rebounded in 2009, then rose further in 2010 and 2011; US Census Bureau, "Nonemployer Statistics," https://www.census.gov/programs-surveys/nonemployer-statistics.html.

support facilities—may have dampened new business formation since they provided support for incumbents and thus deterred exit.[11]

Even though some factors have been more favorable for business formation in the pandemic than in the Great Recession, an open question has been whether genuine employer business creation would result. Historically, likely employer applications have been strongly predictive of actual firm entry, with a national correlation of 0.9 and an elasticity roughly centered on one at the aggregate level, within states, and within industries.[12] But one might fear that the transition rate from applications to actual businesses could change in the pandemic. Perhaps especially in the early months of the pandemic, maybe there was a surge in nascent entrepreneurship—individuals thinking about starting a business—without necessarily making the transition to an actual new business. This is a core question we address by providing available evidence on actual employer business formation below, but first we delve further into the applications themselves.

II.B. Sectoral Patterns of Applications

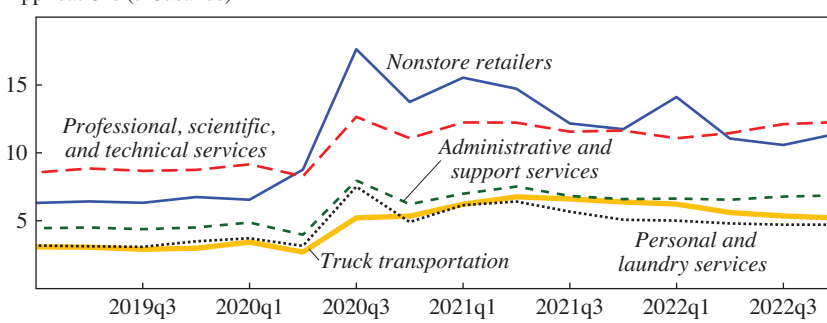
One clue about the economic substance of surging applications is the pattern across industries. For likely employer applications, data are only available at the broad sector level; while interesting (and discussed below), this level of industry detail misses important stories. For more detail, we use total applications, which are available at the three-digit NAICS industry group level (published as a special tabulation after the end of each calendar year—currently these data are available through the end of 2022). We use the total applications series with some caution given our focus on employer business entry, but we note that there has been a coincident surge in likely employer and likely nonemployer applications at observable national, state, and industry levels.[13]

11. There has been some speculation that sole proprietor nonemployer applicants for the Paycheck Protection Program (PPP) had incentives to acquire an EIN to facilitate processing the paperwork requirements of the PPP. This seems unlikely, however; the surge in business formation has persisted long past the last PPP disbursements in mid-2021. Moreover, Breaux and Gurnani (2022) matched PPP and BFS micro data and found that only a very small fraction of PPP applicants applied for an EIN in 2020 and 2021. Only 800 PPP applicants applied for an EIN after they applied for PPP. The average PPP applicant had applied for an EIN about seven years prior to applying for a PPP. This study also rules out the concern that the surge in the BFS in the pandemic reflects any fraudulent PPP applications wherein individuals applied for an EIN to support fraudulent PPP applications.

12. Author calculations on BFS data; for state and industry regressions see online appendix B.

13. At the broad sector level, the correlation in the growth in total applications and likely employer applications (from pre-pandemic to pandemic) is 0.86.

Figure 3. New Business Applications, Selected Three-Digit Industries



Source: Census Bureau Business Formation Statistics.
Note: All applications. Average weekly pace by quarter (seasonally adjusted).

The surge in total applications was highly concentrated among three-digit industries; a Herfindahl-Hirschman index of industry-level applications jumped by more than 10 percent in 2020 versus 2019 and remained historically elevated through 2022 (online appendix figure E5). Indeed, more than 20 percent of the jump in applications from 2019 to 2022 was accounted for by nonstore retailers (NAICS 454), which includes online retail; and more than half of the overall surge was accounted for by just five three-digit industries, shown in figure 3.

The industries making large contributions to overall application growth can plausibly be related to pandemic patterns of work, lifestyle, and business models. Nonstore retailers (NAICS 454) include online retail businesses facilitating shopping from home. Professional, scientific, and technical services (541) is a tech-intensive sector, with about half of its employment in STEM-intensive industries such as architectural, engineering, and related services (5413), computer systems design (5415), and scientific research and development services (5417); business formation in these industries may be related to helping other businesses facilitate pandemic work and lifestyle changes and may also relate to recent technological developments like artificial intelligence (AI).[14] The sector also includes industries such as building inspectors and interior designers potentially associated with the pandemic surge in home sales or rearrangement of home office

14. Many AI-related businesses are classified in this industry; see Library of Congress, "Business Reference Services," https://www.loc.gov/rr/business/BERA/issue31/codes.html. AI firms may also be classified in the Information sector (NAICS 51).

environments. Personal and laundry services (812) include some industries that were likely harmed by the pandemic (e.g., nail salons) but also industries that enhanced work-from-home environments or facilitated pandemic hobbies, such as pet care. Administrative and support services (561) includes employment services that are sometimes important during recessions (e.g., temporary help agencies); industries that may facilitate changes in business models such as document preparation, call centers, and mail carriers; and businesses facilitating work-from-home transitions such as landscaping services and carpet cleaners. Truck transportation (484) includes both general and specialized freight trucking (an example of the latter is NAICS 484210, used household and office goods moving); such businesses likely benefited from changes to the use of commercial real estate, the shift toward online shopping, and the rotation of consumer spending away from services and toward goods.

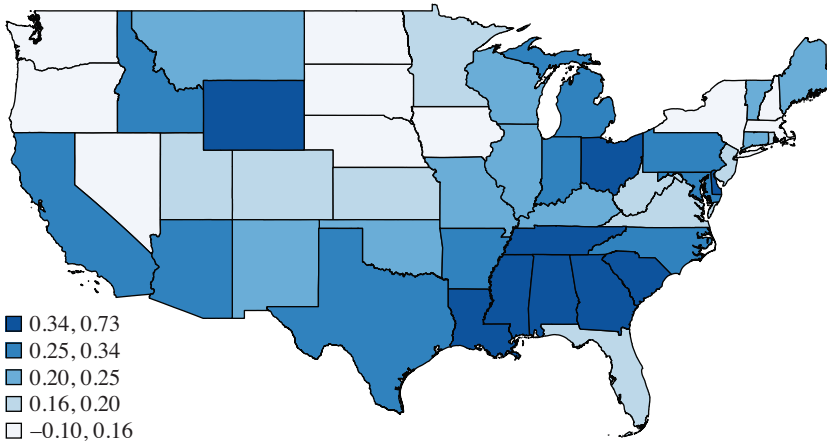
The patterns in figure 3 also hint at interesting changes over the course of the pandemic and its aftermath. Applications for nonstore retailers exhibited the most dramatic surge early in the pandemic; and while this remained elevated at the end of 2022, it has declined substantially from its 2020:Q3 peak. By mid-2022 the highest industry was professional, scientific, and technical services; this tech-intensive industry has exhibited a sustained surge since the beginning, with 2022:Q4 being at about the same pace as 2020:Q3. Truck transportation had a smaller initial surge, peaked in mid-2021, then declined gradually, a pattern consistent with new businesses entering to address supply chain constraints along with the surge in goods consumption, both of which have receded somewhat in recent quarters.

We find similar patterns for likely employer applications at the broad sector level (online appendix figure E2); in particular, we observe strong increases in likely employer applications in the retail trade sector and in "Tech"—a proxy for the high-tech sector that combines professional, scientific, and technical services with the information sector. Interestingly, when we use the projected firm births series from BFS, the two-digit sector that has the highest level of applications during the pandemic is the high-tech sector (online appendix figure E3). As discussed in online appendix B, the predicted start-up series (PBF4Q and PBF8Q) is a better predictor than HBA of actual start-ups, particularly at the sector level.

II.C. Geographic Patterns of Applications

We next analyze spatial variation in applications, and we introduce a simple measure of growth in applications per capita in the pandemic

Figure 4. Growth in Likely Employer Applications per Capita, 2020–2022 versus 2010–2019



Source: Census Bureau Business Formation Statistics and population estimates.
Note: Difference of average (log) likely employer applications per capita, 2020–2022 versus 2010–2019.

relative to the pre-pandemic norm, which we denote as g. We define g as follows, using annual data at various levels of geography:

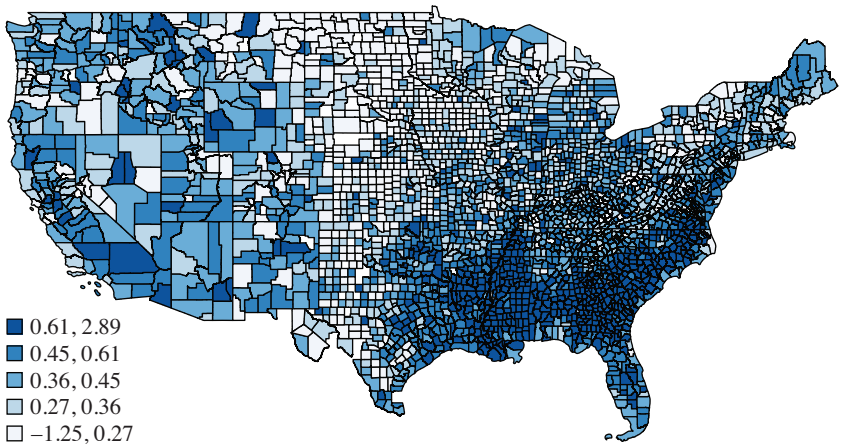
$$g = \frac{1}{3}\sum_{t=2020}^{2022} \ln\left(x_t\right) - \frac{1}{10}\sum_{t=2010}^{2019} \ln\left(x_t\right), \tag{1}$$

where x_t is applications per capita in year t.[15] That is, we study the difference between the average of (log) applications per capita in 2020–2022 and the average of (log) applications per capita during 2010–2019.

Using likely employer applications, figure 4 shows substantial variation across states, with the highest-growth states having growth rates of between 34 and 73 log points while the lowest-growth states exhibit little or no growth. Growth was particularly strong in the South and also parts of the West (e.g., California).

15. In all of our analyses of spatial variation, we focus on per capita variables using Census Bureau county-level population estimates. Karahan, Pugsley, and Şahin (2019) highlight that spatial variation in start-ups is connected to spatial variation in demographic factors such as population growth. Computing measures using annual population estimates helps take this into account, though investigating population migration and its connection to the patterns of start-up dynamics during the pandemic would be of independent interest.

Figure 5. Growth in Total Applications per Capita, 2020–2022 versus 2010–2019



Source: Census Bureau Business Formation Statistics and population estimates.
Note: Difference of average (log) all applications per capita, 2020–2022 versus 2010–2019.

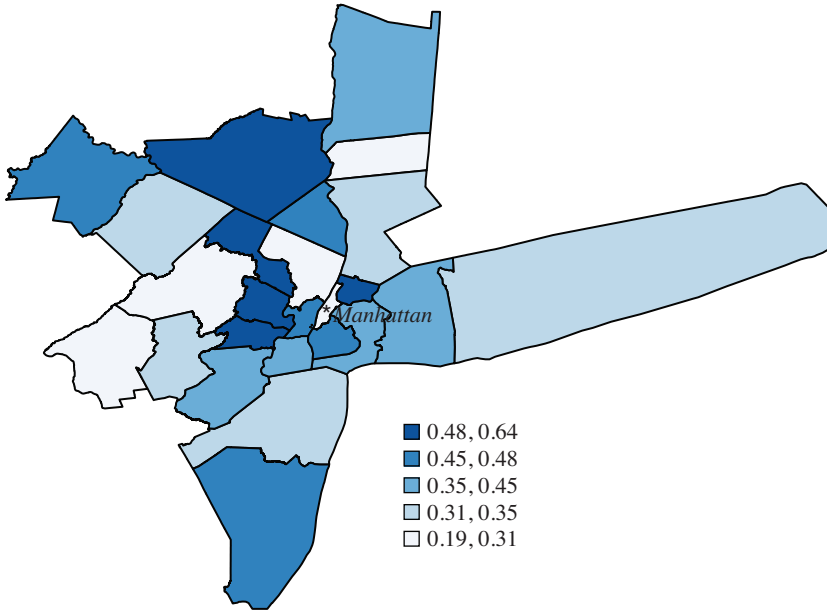
More variation can be seen at the county level, though at this level we must use total applications rather than likely employer applications.[16] Growth in business applications has been widespread across US counties; more than 95 percent of counties saw a higher pace of applications during 2020–2022 than during 2010–2019, on average. Figure 5 provides the county analog to figure 4; the rapid growth in the South is evident in the county map as well, but there are pockets of rapid growth throughout the country.

While a small number of counties actually saw declines in applications per capita, the median county saw an increase of 40 log points, and the highest quintile saw growth of between 61 and 289 log points. The variation in county-level growth suggests material geographic restructuring, with some counties experiencing dramatically more business applications per capita than in pre-pandemic times.

Much of the variation across counties reflects larger geographic shifts: variation between Census Bureau divisions accounts for 25 percent, variation between states accounts for almost 50 percent, and variation between commuting zones accounts for 70 percent of the between-county variation

16. At the state level, the correlation in the growth in total business applications and likely employer applications (pre-pandemic to pandemic) is 0.96.

Figure 6. New York City: Growth in Applications per Capita, 2020–2022 versus 2010–2019



Source: Census Bureau Business Formation Statistics and population estimates.
Note: Difference of average (log) all applications per capita, 2020–2022 versus 2010–2019.

in total application growth (reported in online appendix table F1). However, counties vary considerably in scale, and even though we are examining growth in applications per capita, the latter is increasing in initial county population (and population density). Among counties that are part of large core-based statistical areas (CBSAs), those with population above 1 million, about 50 percent of the between-county variation is accounted for by between-CBSA effects; over half of the US population is in these large CBSAs, so exploring the variation within large CBSAs is of independent interest.

As an example of within-city variation, figure 6 zooms in on the counties of the New York City area (which includes counties in New York State, New Jersey, and Pennsylvania), again reporting growth in (total) applications per capita as calculated in equation (1).

Growth of applications per capita in New York City counties ranges from 19 to 64 log points. We also observe a striking "donut" pattern: growth is stronger outside New York County (i.e., Manhattan—the central business

district of the city) than inside it.[17] These patterns are broadly consistent with zip code–level patterns documented earlier by Fazio and others (2021) using state business registrations; those authors find that, after the widespread initial registration decline early in the pandemic, Manhattan registrations returned to their 2019 pace while the Bronx, Harlem, and parts of Brooklyn saw historic registration growth.[18] Duguid and others (2023) find similar results for retail establishments based on credit card transaction data for the country as a whole; the authors report relatively weak (or negative) establishment growth in core downtown areas, with stronger growth in inner suburbs (though not in outer suburbs).

The donut pattern is apparent in other major cities as well; for example, online appendix figure E7 shows the state of Washington, where King County—the central business district for Seattle—shows less application growth than surrounding counties.[19] In unreported results, we visually observe a similar donut pattern in other cities, in the sense that a number of surrounding (close in and outlying) counties within CBSAs exhibit higher growth in applications per capita than the county that contains the central business district.[20]

The donut pattern we observe for applications appears related to popular pandemic themes about high-density downtown areas and the transition of many workers to work-from-home (WFH) activity. We more formally explore the relationship between the growth of applications, density, and WFH within cities using regressions reported in online appendix table F9. In particular, at the county level we regress growth of total applications per capita on population density, establishment density (from QCEW data), and growth of WFH activity (from ACS data, where the fraction of workers working from home is based on location of residence). We find highly nonlinear, statistically significant empirical relationships for all three covariates. There is an alternating negative linear effect, positive quadratic effect, and negative cubic effect with magnitudes implying the linear negative term dominates for low values (of density and change in WFH share), the positive quadratic becomes relatively more important for larger values,

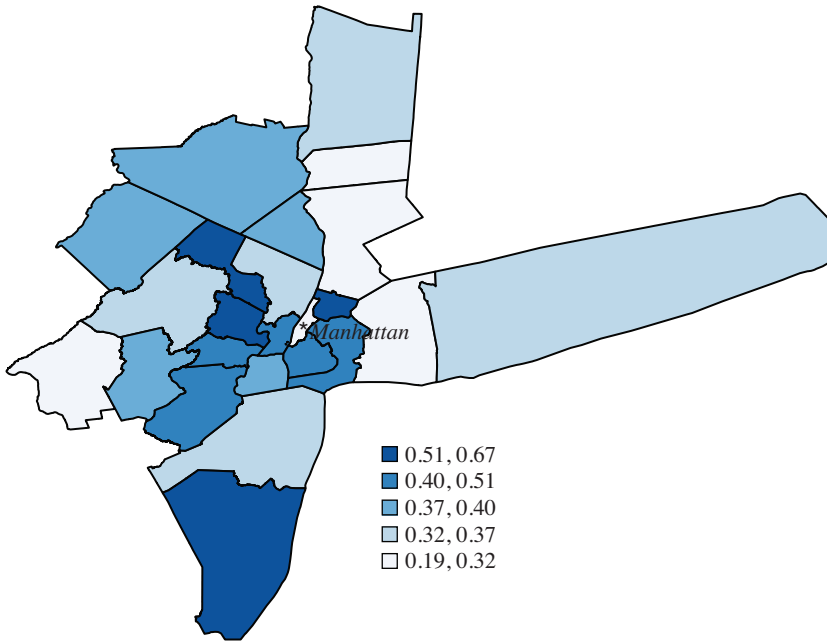
17. Donut-like patterns have been observed on other dimensions such as housing and work, as documented by Ramani and Bloom (2021) among others.

18. Online appendix figure E6 shows that prior to the pandemic Manhattan was one of the top-ranked counties in the NYC CBSA in terms of applications per capita.

19. Online appendix figure E8 shows that prior to the pandemic King County was one of the top-ranked counties in the state of Washington in terms of applications per capita.

20. We hypothesize this effect would be even more prevalent using tract-level data—an approach that awaits the micro data on applications integrated with the LBD.

Figure 7. New York City: Predicted Growth in Applications per Capita, 2020–2022 versus 2010–2019 (Spatial Model)



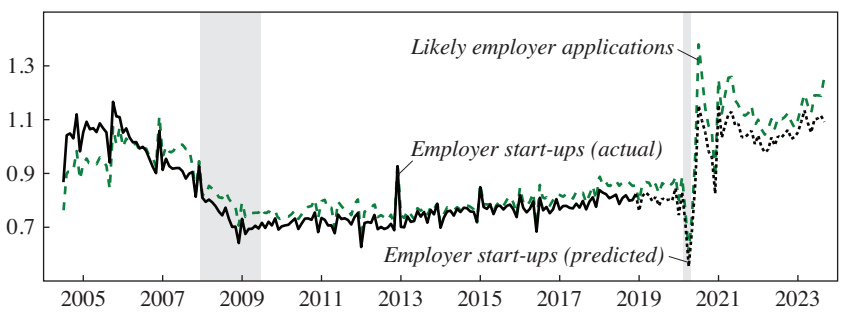
Source: Census Bureau Business Formation Statistics and population estimates; author modeling.
Note: Predicted difference of average (log) all applications per capita, 2020–2022 versus 2010– 2019.

then the negative term kicks in for very large values. In considering these patterns, it is useful to observe that within New York City, Manhattan has the highest population density and establishment density and a mid-range growth of WFH.[21]

We also consider a more complex spatial regression specification where we include a cubic of all of these terms for both own and adjacent counties (online appendix table F10). We find that each of these covariates have significant own- and adjacent-county effects in this multivariate specification. Given the complexity of this specification, we focus on the overall predictive power; the R^2 of this specification is 0.77, compared to 0.49 in a specification with only CBSA fixed effects. Figure 7 shows that the predicted

21. Fazio, Guzman, and Stern (2020) observe a positive, but not statistically significant, linear relationship between density and business registration growth in their eight-state sample, though they do not study nonlinear dimensions. A nonlinear relationship is consistent with Duguid and others (2023), who also find nuanced relationships with WFH activity.

Figure 8. High-Propensity Business Applications and Start-ups Eight Quarters Ahead



Source: Census Bureau Business Formation Statistics.

Note: Start-ups within eight quarters. Seasonally adjusted. Normalized by average 2006 levels. Shaded areas indicate NBER recession dates.

variation in counties in the New York City CBSA from this spatial model closely corresponds to the actual application pattern (compare to figure 6). Put simply, we are able to approximately replicate the within–New York City donut pattern using population density, establishment density, and growth of WFH activity in own and adjacent counties—consistent with the broader high model fit for all cities suggested by the R^2 of 0.77.

II.D. Applications and Actual Firm Births in the BFS

There is typically a lag between EIN application and new employer firm entry, even conditional on a successful transition. In much of our analysis of employer entry from other sources we use quarterly data, annual data, or growth rates based on the difference between pre-pandemic and pandemic averages, mitigating this lag.[22] Figure 8 shows a tight relationship between likely employer applications and employer firm births within eight quarters. The solid line provides an index of actual employer firm births within eight quarters (through 2018:Q4), and the dotted line is an index of projected employer firm births within eight quarters. The surge in likely employer applications in the pandemic is accompanied by a surge in projected business formations.[23]

22. Dinlersoz and others (2023) show that in the same quarter as the application, the historical transition rate of applications with planned wages has been about 14 percent; the transition rate is 35 percent after four quarters and 40 percent after eight quarters.

23. In interpreting this finding, it is important to emphasize that the projected series takes into account the full range of application characteristics. We further discuss the relationship between the likely employer series and the projected firm birth series in online appendix B.

Online appendix tables F2 and F3 provide more detail about this tight relationship between applications and actual employer start-ups. As a rough approximation, the (pre-pandemic) elasticity of new employer firms within eight quarters with respect to likely employer applications is centered on one in both the aggregate time series and in state-by-time pooled data. These historical relationships as well as the projected series suggest strongly elevated employer entry during the pandemic as well—but with some lag relative to the timing of applications. The lag between application and new employer entry was increasing prior to the pandemic (see online appendix figure E4). While the actual transitions are not yet available beyond 2020, we explore these relationships below using a variety of available employer entry rates.

III. New Employer Businesses in the Pandemic

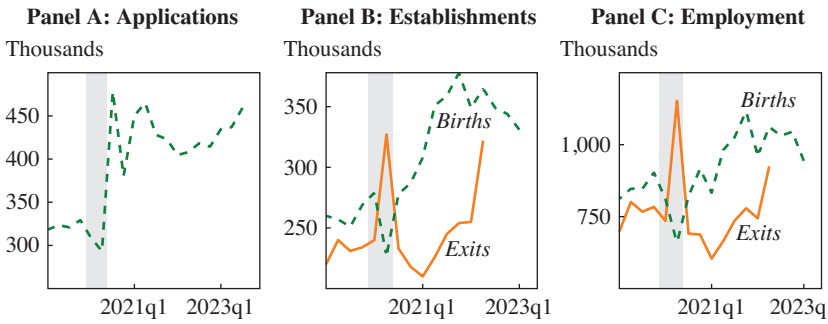
We now turn to data on actual employer business formations during the pandemic, expanding on the data first shown in figure 2. Here we draw on several sources: we use BED quarterly establishment births and openings data through 2023:Q1, BED annual firm births data through March 2022, and QCEW quarterly net establishment births data through 2023:Q1 (which permit finer geographic and industry detail than BED data). Importantly, the gold standard data set for tracking true employer firm births is the Census Bureau's BDS, which features a more comprehensive firm identifier than the BED (see discussion in online appendix A); we report two different BDS series in figure 2, but these data do not currently cover a significant portion of the pandemic period.

The BED and QCEW have the key advantage of timeliness, though the most timely data are on establishment entry (gross entry in BED and net entry in QCEW), which include not only new firms but also new establishments of incumbent firms (e.g., new Starbucks locations). While our primary focus is on new firms, new establishments opened as expansions of existing firms are of independent interest, since such establishments are important components of the reallocation of activity across business locations. Moreover, it is likely that new establishments of existing businesses reflect similar incentives of new firms to take advantage of the market opportunities that arose in the pandemic and its aftermath.

III.A. Aggregate Establishment and Firm Entry: Gross and Net

Figure 9 shows quarterly data on high-propensity business applications (panel A), BED establishment births and exits (panel B), and jobs created (destroyed) by births (exits) (panel C).

Figure 9. Business Applications, Establishment Births, and Exits



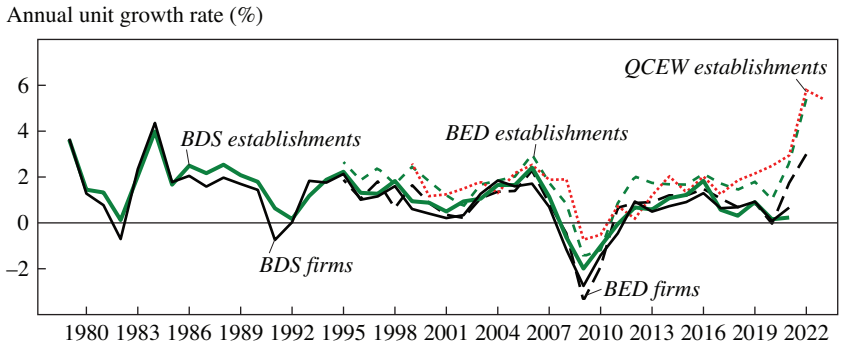
Source: Census Bureau Business Formation Statistics (BFS) and BLS Business Employment Dynamics.
Note: Seasonally adjusted. Shaded areas indicate NBER recession dates. High-propensity applications. Panel C shows jobs created by births and jobs destroyed by exits.

The surge in establishment births is especially pronounced starting in 2021:Q2—several quarters after the initial surge in applications in July 2020 but also after the second wave of the surge in applications in early 2021. It is not surprising that there is some lag since, as discussed above, it can take up to eight quarters for applications to transit to employer businesses—conditional on transiting at all. Like business applications, establishment births have reached record levels during the pandemic. Note also that births have been well in excess of exits, aside from the initial exit surge in 2020:Q2.

As shown in panel C, job creation from establishment births has been above one million per quarter, on average, during 2021:Q2–2023:Q1—a historically high pace. Establishment birth has played a significant role in the pandemic job recovery, accounting for more than 10 percent of gross private job creation from 2020:Q3 through 2023:Q1; at a quarterly frequency, in 2022:Q4 establishment births' share of gross job creation reached 12.9 percent for the first time since 2007. While this increase in job creation from births is striking, the surge in the number of establishment births (panel B) is proportionally greater than the surge in birth employment; the average size of a new establishment birth declined from about 3.3 jobs in 2019 to 2.9 jobs in 2022.[24] As we discuss in section V.B below, average firm entrant size also stepped down in the pandemic—though incumbent size declined as well.

24. Author calculations from BED data.

Figure 10. Net Growth of Establishments and Firms



Source: BDS; BED; and QCEW.
Note: Annual Davis-Haltiwanger-Schuh denominator (DHS) growth rate of unit counts, first quarter versus one year earlier.

The elasticity of establishment births with respect to likely employer applications has, if anything, strengthened in the pandemic—at least at the national level; we obtain this evidence with simple regressions of establishment births per capita on applications (online appendix table F2). For the aggregate series we actually find a higher elasticity of establishment births when we include the pandemic period than if we end the sample in 2019. Online appendix table F4 reports state-by-quarter regressions and table F6 reports sector-by-quarter results, in which the pre-pandemic elasticities (shown on the top panel of each table) are generally similar to those estimated on pandemic-inclusive data (bottom panel). We also examine the relationships between firm births, establishment births, and the projected start-up series from the BFS. Online appendix tables F7 and F8 illustrate three findings. First, there is a strong positive historical (pre-pandemic) relationship between BFS predicted firm births and actual firm and establishment births. Second, this relationship remains strong during the pandemic for establishment births. Third, especially for the sector-based results, the elasticities are substantially higher using the BFS projected start-ups series compared to those using the likely employer applications. We discuss these analyses more in online appendix B.

It is clear from the BED data in figure 9 that net establishment entry surged in the pandemic; this fact is corroborated in other data sources and for firms as well. Figure 10 shows annual net growth of firm and establishment counts from the BDS, BED, and QCEW. Reassuringly, the various series track each other well through March 2020, after which the

BDS becomes unavailable. Net establishment growth was strong in 2021 and, especially, 2022 and 2023.[25] Firm growth was similarly impressive, as the total number of firms (in BED data) increased by more than 250,000 from March 2020 through March 2022, from under 5.3 million to more than 5.5 million. The largest surge is from March 2021 to March 2022—broadly consistent with the finding that the increase in establishment births is especially pronounced starting in 2021:Q2. In online appendix figure E9 we report similar results if growth is calculated on a per capita basis.

Here we have focused on true establishment birth and exit; temporary closings and reopenings of establishments also played a large role in early pandemic labor market dynamics. In online appendix figure E12 we report total establishment openings and closings, and figure E13 shows reopenings (i.e., openings minus births) and temporary closings (i.e., closings minus exits). In 2020:Q2, more than 400,000 establishments closed temporarily, with nearly 1.8 million associated jobs. Reopenings jumped in the following quarter, accounting for 1.2 million jobs in 2020:Q3 and nearly 800,000 jobs in 2020:Q4. These patterns imply a need for caution in the use of establishment openings out of context—especially in 2020:Q3; the patterns also highlight the large role of temporary job dislocation in the early pandemic labor market.

While establishment reopening and temporary closure was a significant feature of the pandemic—particularly in early quarters—the cumulative job reallocation associated with births and exits is even a bit larger.[26] Over the 2020:Q2–2023:Q1 period, job reallocation from establishment births and exits cumulated to 20.6 million jobs, with births contributing 11.4 million and exits 9.2 million. Reallocation from births and exits necessarily reflects permanent job reallocation. During the same period, temporary closings and reopenings cumulated to 17.5 million jobs, with temporary closings contributing 9.1 million and reopenings 8.4 million. In contrast to births and deaths, these job flows associated with temporary closings and reopenings reflect transitory reallocation—although it may be that some workers who lost

25. We are not the only researchers to notice the striking surge in establishment counts; for example, O'Brien (2022) highlights the net growth of establishments and explores cross-city variation.

26. For the calculations in this paragraph, we impute job destruction from establishment exit in quarters after exit data end (that is, after 2022:Q2) by setting exit job destruction equal to its average over 2020:Q3–2022:Q2, which is a bit over 700,000 jobs per quarter. We use this imputed exit job destruction path to estimate employment associated with temporary establishment closures in quarters for which exit data are unavailable (but total closure employment data are available).

their jobs to temporary closings did not return to the same employer, since reopenings took some time. We discuss the implications of these dynamics for job and worker reallocation further below.

III.B. Sectoral Patterns of Employer Business Entry

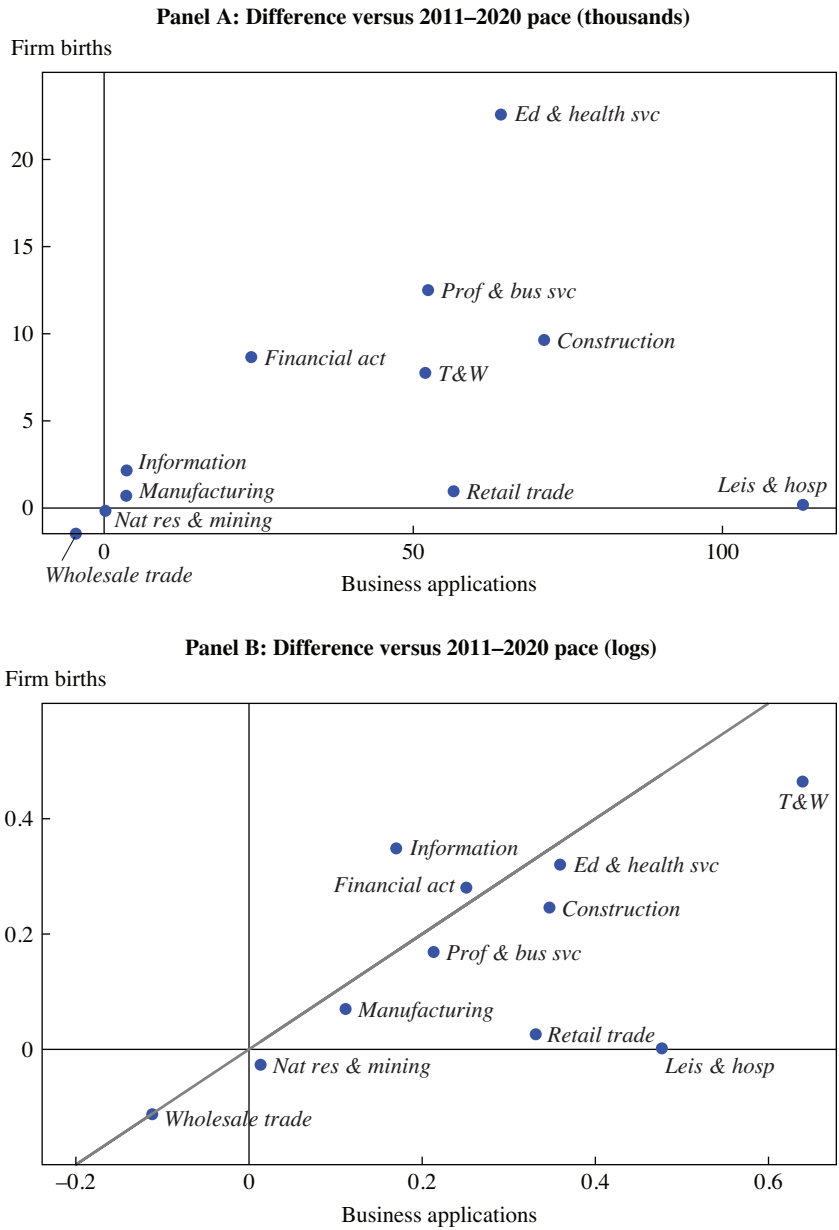
As noted in section II.B, the industry pattern of business applications is consistent with broader economic restructuring in the pandemic. We next ask whether these industry patterns are reflected in data on actual employer business formation. Annual firm births by broad sector are available from the BED through March 2022; the scatterplots in figure 11 compare pandemic firm births with likely employer applications by sector, where we focus on pandemic growth relative to pre-pandemic norms as described in equation (1).

Panel A in figure 11 gives insight into the contribution of different sectors to the aggregate surge in firm births and likely employer applications by measuring the average level of births or applications—in thousands—during the pandemic versus the pre-pandemic pace. Educational and health services, professional and business services, and construction are sectors with large increases in both firm births and likely employer applications, accounting for a large share of the aggregate surges in both.

Panel B is more informative about growth within sectors, as it is based on the log difference between pandemic and pre-pandemic norms. Sector-level growth in firm births and business applications is strongly positively related, with most sectors lining up reasonably close to the 45-degree line. Transportation and warehousing, information, education and health services, financial services, construction, and professional and business services are all sectors with large growth (approximately 20 percent or larger) of both applications and firm births. The retail trade sector is notable, however, for having a smaller surge in firm births than in applications; this could reflect the differing nature of the 2020 application surge (which, as discussed above, was led by online retail) versus the later pandemic surge, where other sectors became more important.[27] It may be that the early surge in applications, especially in sectors like online retail, saw lower rates of transition to employer business formation. Indeed, the BFS itself suggests this; in online appendix figure E11, we find that the sector-level relationship

27. More industry detail can be seen in online appendix figure E10, which narrows down to the three-digit NAICS level (but necessarily relies on establishment openings and total applications). We find strong, statistically significant relationships using this detailed variation.

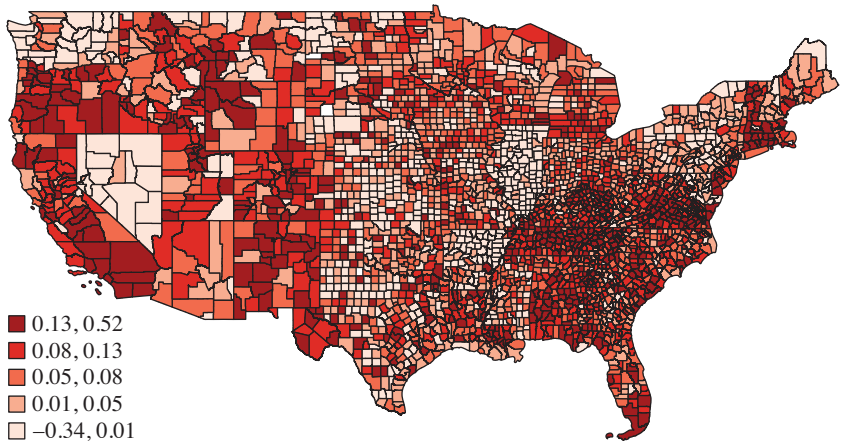
Figure 11. Firm Births and Business Applications, Industry Detail



Source: Business Employment Dynamics (BED) and Business Formation Statistics (BFS).

Note: Average pace during 2021–2022 versus average pace during 2011–2020. Panel A expressed as average annual pace. Solid line is the 45-degree line. "T&W" is transportation and warehousing. Years end in March. High-propensity applications.

Figure 12. Net Establishment Growth from Pre-pandemic to Pandemic



Source: QCEW and Census Bureau population estimates.
Note: Difference of average (log) establishments per capita, 2020–2022 versus 2010–2019.

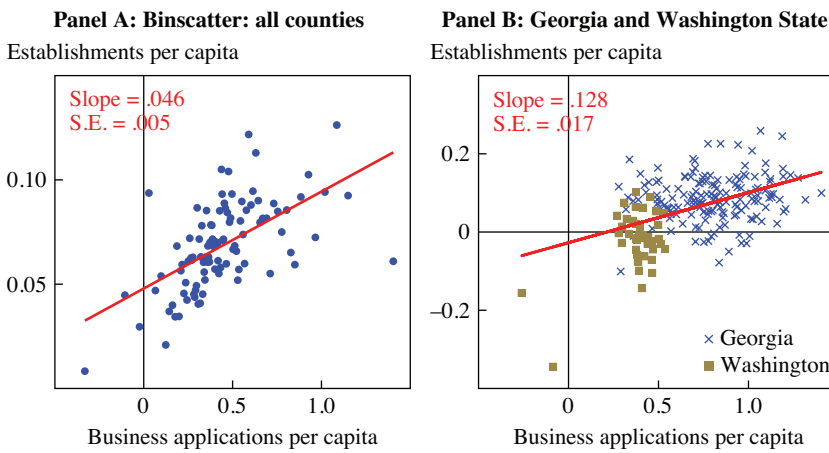
between firm births and BFS projected firm births is even stronger, with the retail sector being less of an outlier.

III.C. Geographic Patterns of Employer Business Entry

Given the striking geographic pattern of business applications described in section II.C, we next explore county-level correlations. Data limitations continue to bind, however, as BED establishment birth (or opening) data are not available at the county level, so we focus on net establishment entry (i.e., change in the number of establishments) in QCEW data. To start, we consider the spatial variation in growth of establishments per capita between the pre-pandemic and pandemic periods using the same measure as implied by equation (1). Figure 12 highlights substantial variation in the growth of establishments per capita across counties (this figure can be compared usefully with figure 5, the analogous map for business applications). In the top quintile, establishments per capita increased between 13 and 52 log points while in the bottom quintile establishments per capita declined.

The spatial patterns in figures 12 and 5 are broadly similar, with the South and parts of the West standing out as having especially high growth in both applications per capita and establishments per capita. We can see this more formally in panel A of figure 13, which is a binscatter relating county-level growth in total establishments per capita to growth in applications per capita, 2020–2022 versus 2010–2019, following equation (1).

Figure 13. Net Establishments Growth versus Applications Growth



Source: QCEW and BFS.

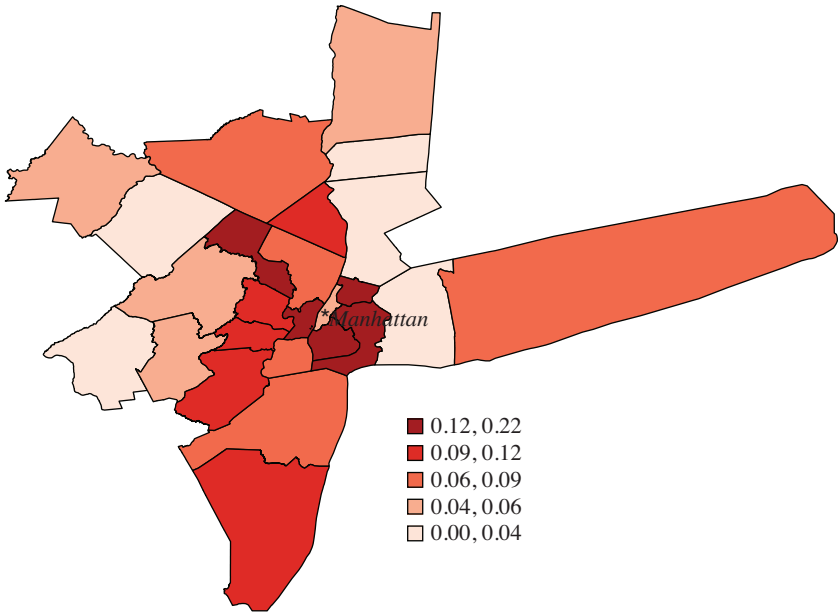
Note: County-level log differences of 2020–2022 versus 2010–2019 levels. Straight line is a regression line with reported slope and standard error. Total applications. Panel A is a binscatter with one hundred bins.

We observe a tight, highly statistically significant relationship between establishment growth and applications. Of course, net establishment growth conflates establishment birth and exit, and the latter has likely been an important margin of local economic adjustment during the pandemic period; see Decker and Haltiwanger (2022) and Crane and others (2022) for discussion (though recall that figure 9 shows that establishment death was not materially elevated after its initial spike in 2020:Q2, with the exception of 2022:Q2).[28] Moreover, as in our three-digit industry scatterplots above, at the county level we have total business applications, not the narrower category of high-propensity applications, though recall that total applications and high-propensity applications have moved together in the pandemic. The strong spatial relationship between net establishment entry and total applications suggests that surging business applications are related to growth in net entry in the geographic cross section.[29]

28. The 2022:Q2 establishment exit jump is puzzling, and we confirmed with BLS staff that it is not an artifact of any obvious measurement or scope issue. We note, however, that exits are measured with a lag, and that parts of the data used to measure exit in this quarter could still be revised in future years.

29. The small slope coefficient reflects the much greater variation in the growth of applications per capita relative to growth of establishments per capita, which is apparent from the chart axes.

Figure 14. Net Establishment Growth, New York City



Source: QCEW and Census Bureau population estimates.
Note: Difference of average (log) establishments per capita, 2020–2022 versus 2010–2019.

We provide some concrete perspective into our county maps and the binscatter just mentioned by focusing on the counties in two states: Georgia and Washington. Panel B of figure 13 depicts the growth in applications and establishments for counties in just these two states; Georgia (crosses) is a state with high growth on both margins, while Washington (squares) is not. Interestingly, this between-state pattern holds pervasively across counties within these respective states.

As another specific example, figure 14 shows net establishment growth for counties of New York City in the same manner as figure 6. While not identical to the pattern of application growth, we still observe a donut pattern of strong growth in establishments per capita in the city suburbs, with less growth in the city center of Manhattan.

We provide further perspective on these geographic patterns in online appendix C. We find, for example, that the high-growth counties in terms of net establishment growth in the NYC area have higher growth rates than Manhattan across a wide variety of industry sectors. Some of this

reflects sectors that are apparently supporting the change in the habits of the daytime population (e.g., large increases in sectors such as leisure and hospitality—NAICS codes 71 and 72). However, we also observe the high-growth counties having higher growth in high-tech sectors like information (51) and professional, scientific, and technical services (54). Similar observations apply to high-growth states such as Georgia relative to low-growth states such as Washington.

Our geographic exercises, like our industry exercises, suggest a strong relationship between business applications and actual employer business growth. Moreover, these patterns are consistent with thriving business creation in industries that complement pandemic changes in work and lifestyles as well as movement of some forms of economic activity from city centers to outer areas. Notably, our geographic analysis is all done on a per capita basis, so these flows of businesses do not simply reflect underlying population flows.

IV. Worker Flows and Business Formation

The pandemic labor market has featured several notable patterns, including mass layoffs followed by rapid job growth, migration, and a large number of workers quitting their jobs (which has been called the Great Resignation). A natural question is whether these labor market patterns have any relation to the surge in business formation. In section III.A we described the significant role of firm and establishment birth in gross and net job growth in the pandemic; and in section III.C we reported striking geographic patterns consistent with popular stories about migration flows (north to south, inner cities to outer cores) but that reflect flows above and beyond simple population moves.

In this section, we focus specifically on quits and layoffs or, where necessary, close proxies for quits and layoffs. The early pandemic period was characterized by a massive spike in layoffs; while many of these proved temporary (Cajner and others 2020), the 2020:Q2 spike in establishment deaths (figure 9) indicates that there was also considerable permanent job destruction. Separately, the pace of quits rose to record levels—and well above its pre-pandemic trend—in late 2021 and early 2022.

Workers who experience a permanent separation through either quits or layoffs could be joining a new business either as the entrepreneur or as an early employee. Indeed, since quits are thought to be dominated by job-to-job flows, workers who quit likely had a job to go to at the time of the

quit.[30] But the administrative micro data required to track these flows on a comprehensive basis are not yet available. Instead, we examine patterns at the aggregate and spatial levels as we have in previous sections.

For this purpose, we exploit data from the Census Bureau's Quarterly Workforce Indicators (QWI) and other sources. The QWI provide information on hires (i.e., new worker-firm matches), separations (broken worker-firm matches), job creation (growth in firm employment), and job destruction (contraction of firm employment) in various granular tabulations.[31] We take advantage of that granularity to decompose separations into job destruction and what we denote—following Davis and Haltiwanger (1992)—as excess separations (the difference between separations and job destruction).

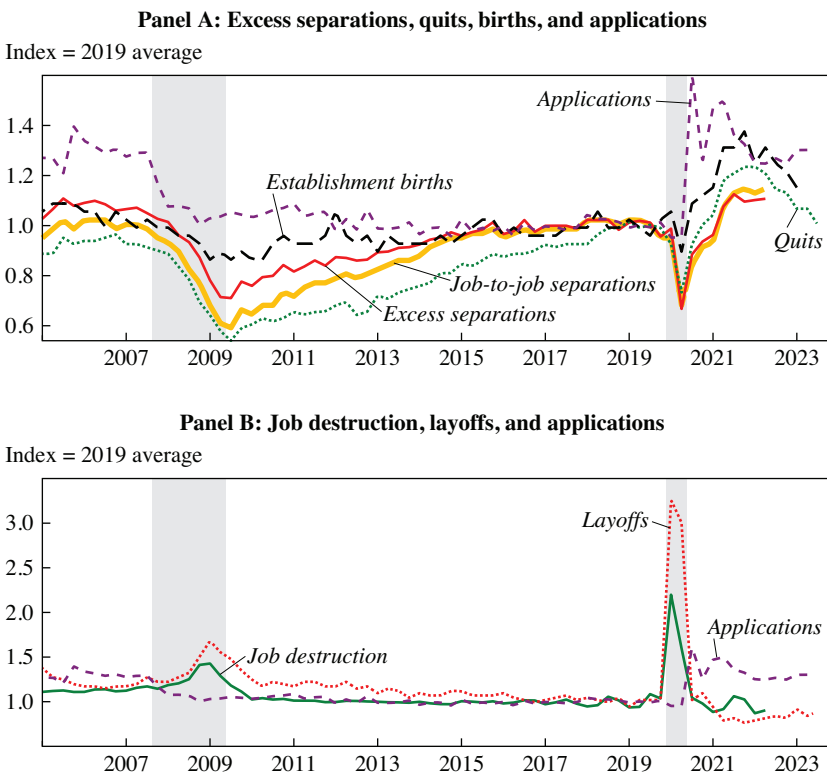
It is important to grasp the intuition of excess separations. Separations include both layoffs and quits. Workers may be separated from jobs because those jobs are being destroyed as a firm contracts; for example, a firm may be eliminating a position entirely as part of a downsizing or restructuring plan. In these cases, there is no excess separation, and worker and job flows are equal. But many workers are separated from jobs while those jobs continue to exist and will be filled by another worker. A likely reason for such a separation is that the worker is quitting the job to start a new job elsewhere. Both conceptually and historically, job destruction and layoffs track each other well, and excess separations and quits track each other well (Davis, Faberman, and Haltiwanger 2012).

Figure 15 reports worker flows (i.e., quits and layoffs and their proxies), establishment births, and business applications. Panel A shows excess separations from QWI and the standard quits series from the BLS Job Openings and Labor Turnover Survey (JOLTS), along with BED establishment births and BFS high-propensity business applications (all series indexed to 2019 rates). Prior to the pandemic, quits and excess separations moved in similar patterns (albeit with some level shift), consistent with their close conceptual relationship. This co-movement continued in the pandemic, with an initial drop in quits and excess separations followed by a recovery to historic levels (admittedly more dramatic for quits). Over the same period, business applications and actual establishment births surged as well. Panel A shows one other series as well: job-to-job separations from

30. Elsby, Hobijn, and Şahin (2010) find that layoffs, not quits, account for cyclical flows from employment into unemployment. Davis, Faberman, and Haltiwanger (2012) find a tight connection between job destruction and layoffs, and job-to-job flows are tightly linked with quits; see Molloy and others (2016), including the comment by Haltiwanger (2016).

31. See online appendix A for detail about the QWI and how we use it.

Figure 15. Worker Flows and Applications



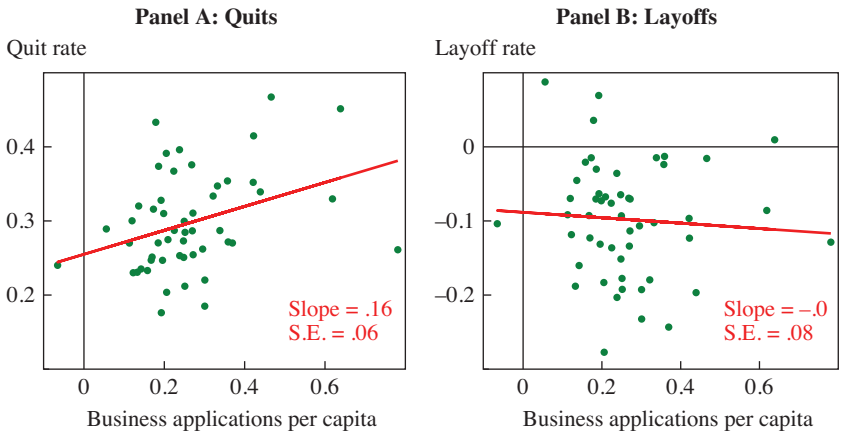
Source: QWI; JOLTS; BED; BFS; and the US Census Bureau J2J.

Note: Index of series expressed relative to employment or, for births, to establishments; seasonally adjusted. Applications are likely employers (HBA). Shaded areas indicate NBER recession dates.

the Census Bureau's Job-to-Job Flows (J2J), which is closely related to the QWI. This series measures separations of workers in which the worker quickly starts a new job with a different firm; as suggested by the discussion, excess separations closely track job-to-job separations in figure 15, as both are closely related to quits.

Panel B of figure 15 shows the spike in job destruction and layoffs in the second quarter of 2020. Both spikes are short-lived and, as noted previously, the layoffs in particular reflect a surge in temporary layoffs. Using data from the Current Population Survey (CPS) on inflows to unemployment from employment (using those entering unemployment in a month based upon duration data), about 85 percent of the massive surge

Figure 16. Quits, Layoffs, and Applications, 2020–2023 versus 2010–2019



Source: JOLTS and Business Formation Statistics (BFS).

Note: State-level log differences of 2020–2023 versus 2010–2019 seasonally adjusted pace. The straight line is a regression line with reported slope and standard error. Data through August 2023.

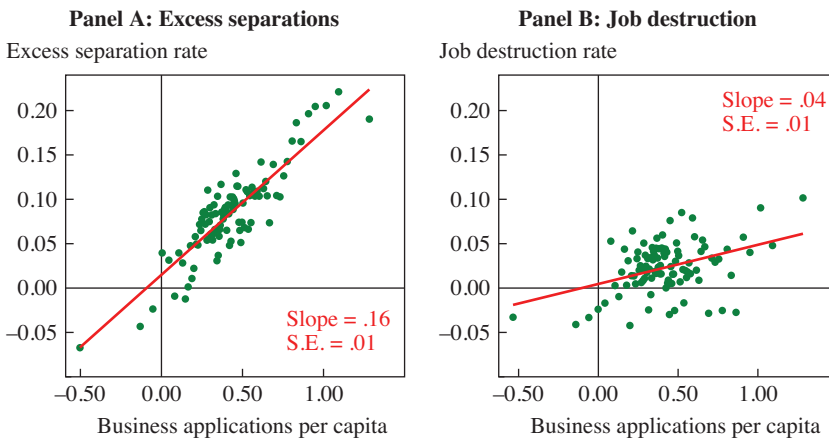
in unemployment inflows in 2020:Q2 was due to temporary layoffs (see online appendix figure E17). Both series drop to low levels after mid-2020, even while business applications surged.

The two panels of figure 15, taken together, are suggestive of a relationship between quits (or their proxy, excess separations) and business formation, consistent with a theory in which workers quit their jobs to start, or join, new businesses. On the other hand, such a relationship between layoffs and business formation is not obviously apparent, as if the surge in business creation does not simply reflect laid-off workers starting businesses due to weak labor market opportunities. Still, these are simply aggregate series.

We therefore turn to spatial variation. We start at the state level, where JOLTS data on quits and layoffs as well as BFS likely employer applications are available; we employ the same approach as prior analyses to study the pandemic relative to pre-pandemic norms. As shown in panel A of figure 16, states with especially large surges in likely employer applications also saw especially large surges in quits during the 2020–2023 period; while there is much variation in both series, there is a substantive positive relationship that is statistically significant.[32] As seen in panel B, there is

32. We apply equation (1) using monthly data for this purpose, computing the mean of the log of series per capita for the pre-pandemic (2010–2019) and pandemic (2020–2023) periods.

Figure 17. Excess Separations, Layoffs, and Applications, 2020–2022 versus 2010–2019



Source: Quarterly Workforce Indicators (QWI) and Business Formation Statistics (BFS).
Note: County-level log differences of 2020–2022:Q2 versus 2010–2019 seasonally adjusted pace. The straight line is a regression line with reported slope and standard error. Binscatter with one hundred bins.

no apparent association between layoffs and high-propensity applications across states, consistent with the aggregate data in figure 15.

We next drill down to the county level, where we can examine related patterns using excess separations and job destruction from the QWI (our proxies for quits and layoffs) and total applications from the BFS. In figure 17, panel A shows a binscatter of county-level growth in the excess separations rate and county-level growth in (total) business applications per capita, where growth is again constructed as in equation (1). We observe a tight, statistically significant spatial relationship between growth in excess separations and growth in business applications. In panel B, though, we observe a much weaker (albeit positive) relationship between job destruction and applications.

While we might imagine multiple mechanisms underlying the observed spatial relationships, one possible explanation is that surging business creation and resulting labor demand is an important component of the overall story of worker flows in the pandemic, including quits. New businesses aggressively poach workers from other firms (Haltiwanger and others 2018) and, therefore, likely contributed to the pandemic reallocation of workers by providing new opportunities in pandemic-friendly industries. We know from figure 9 that job creation by establishment births during 2021 was substantial; with new establishments creating roughly one million

jobs per quarter, some job-to-job flows—arising from excess separations—would likely result.

Interestingly, within cities we find a donut pattern of excess separation growth similar to the pattern for applications (and net establishment births); online appendix figure E19 shows that county-level growth in excess separations for New York City has been greater in the counties surrounding Manhattan than in Manhattan itself.

V. Business Dynamism Revived?

A large body of literature explores declining business dynamism, or the slowing of job and business flows in recent decades, including a decline in the firm entry rate and the share of activity accounted for by young and small firms. The evidence above suggests that the pandemic has been a period of increased dynamism relative to the 2010–2019 period. In this section, we consider the possibility of a return of the higher dynamism pace of the past (pre-2000). While we find noteworthy evidence of substantial economic restructuring during the pandemic—including reallocation of jobs and changes in the firm age and size distribution—we conclude that more time (and data) is needed for a material reversal of pre-pandemic trends.

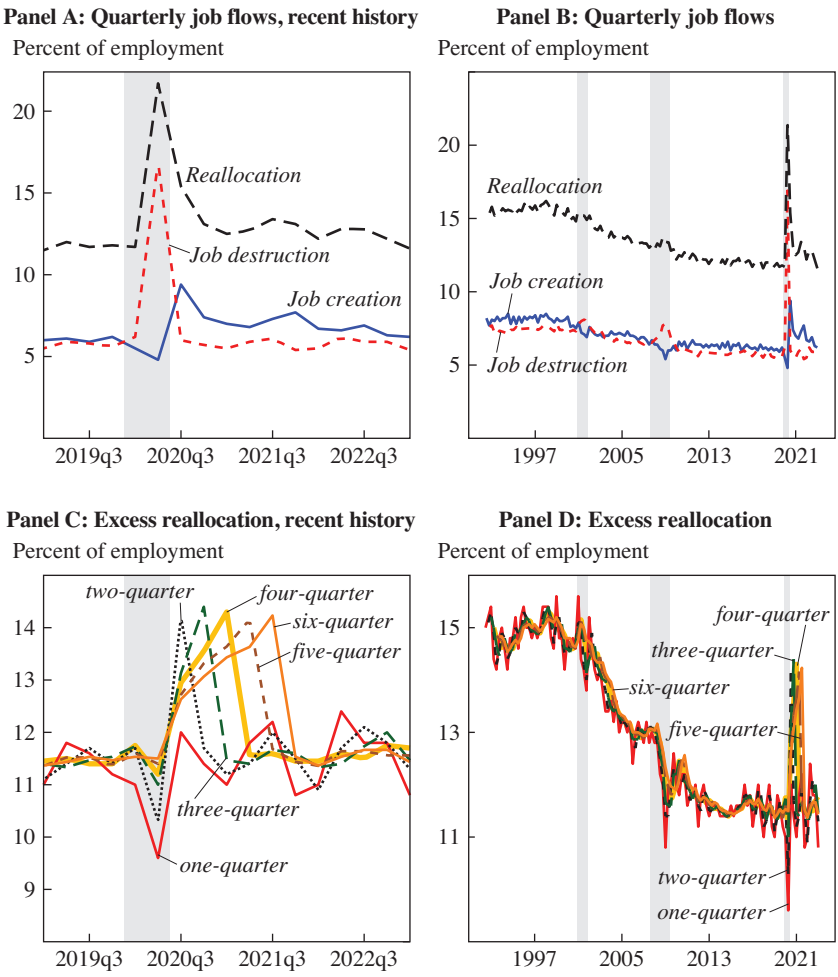
V.A. Job Reallocation

Following literature that goes back a long way (Davis and Haltiwanger 1992), we define the job reallocation rate as:

$$jr_t = \frac{jc_t + jd_t}{\frac{1}{2}\left(e_{t-1} + e_t\right)} \tag{2}$$

where jc_t is gross job creation (total jobs created by entering and expanding establishments), jd_t is gross job destruction (total jobs destroyed by downsizing and exiting establishments), e_t is employment, and t indexes time (quarters, for our purposes). Job reallocation is a summary measure of the reallocation of jobs across expanding, opening, contracting, and closing establishments and is often used as a measure of business dynamism. The denominator in equation (2) is the Davis-Haltiwanger-Schuh (DHS) denominator after Davis, Haltiwanger, and Schuh (1996). Panels A and B of figure 18 show gross job creation, gross job destruction, and job reallocation; panel A zooms in on the pandemic period, while panel B shows a longer view.

Figure 18. Perspectives on Job Reallocation



Source: Business Employment Dynamics (BED).

Note: Reallocation is $jc + jd$, from equation (2). Excess reallocation is $jc + jd - |jc - jd|$, with jc and jd averaged over indicated horizon. Seasonally adjusted. Shaded areas indicate NBER recession dates.

As has been extensively documented in the literature, job reallocation exhibits a downward trend over the last few decades and especially since the early 2000s. More recently, job reallocation spiked early in the pandemic; as shown in panel D, the pandemic spike was historic. The 2020:Q2 spike in reallocation was driven by the surge of job destruction. In the following quarter, reallocation moved down some but remained elevated;

initially this reflected the surge of job creation as temporarily destroyed jobs returned.

There are two critical points to make about the early pandemic spike in reallocation. First, as just noted, the 2020:Q2 spike was driven entirely by surging job destruction and therefore simply reflects net (negative) job growth in that quarter rather than a dynamism phenomenon of simultaneous job creation and destruction across establishments; the 2022:Q3 elevation is similar but driven by job creation. Second, the pandemic was peculiar in that many of the jobs created in 2020:Q3 (and the immediately following quarters) were the same jobs—in the same establishments—that had been destroyed in 2020:Q2, as pandemic business restrictions or voluntary social distancing causing initial business closures and temporary layoffs were followed by quick resumption of business activities and recalls (Cajner and others 2020). As a result, quarterly excess job reallocation (job reallocation in excess of absolute net employment growth, or $jr_t - |jc_t - jd_t|$) actually moved down in 2020:Q2 and has not generally been significantly elevated during the pandemic (this can be seen in the one-quarter line in panel C of the figure, which we discuss more below).

Readers should carefully note that excess reallocation measures can be misleading in quarterly data, as noted in Davis and Haltiwanger (1992) and related work, especially when creation and destruction are decoupled or staggered in terms of timing. A clearer perspective emerges from measuring excess job reallocation using multi-quarter averages of job creation and destruction. Excess reallocation measured at two-, four-, or six-quarter horizons did indeed surge to a pace not seen in more than a decade, as can be seen in panels C and D of figure 18 (which also shows the dip in one-quarter excess reallocation).[33] Excess reallocation measured at multi-quarter horizons (e.g., the six-quarter line in figure 18) was elevated for an extended period in the pandemic, though it came down again in 2022.

Without access to the micro data, we still cannot be certain that this multi-quarter horizon increase in excess job reallocation does not simply reflect job destruction in one quarter followed by job creation in the same establishment in subsequent quarters. To explore this question more, we return to the rich QWI data and focus on between-cell excess job

33. Excess reallocation measured at an h-quarter horizon is given by:

$$er_t^h = \bar{jc}_t^h + \bar{jd}_t^h - \left|\bar{jc}_t^h - \bar{jd}_t^h\right|,$$

where $\bar{jc}_t^h$ is average quarterly job creation over the h quarters leading up to (and including) t, and $\bar{jd}_t^h$ is the corresponding average of job destruction.

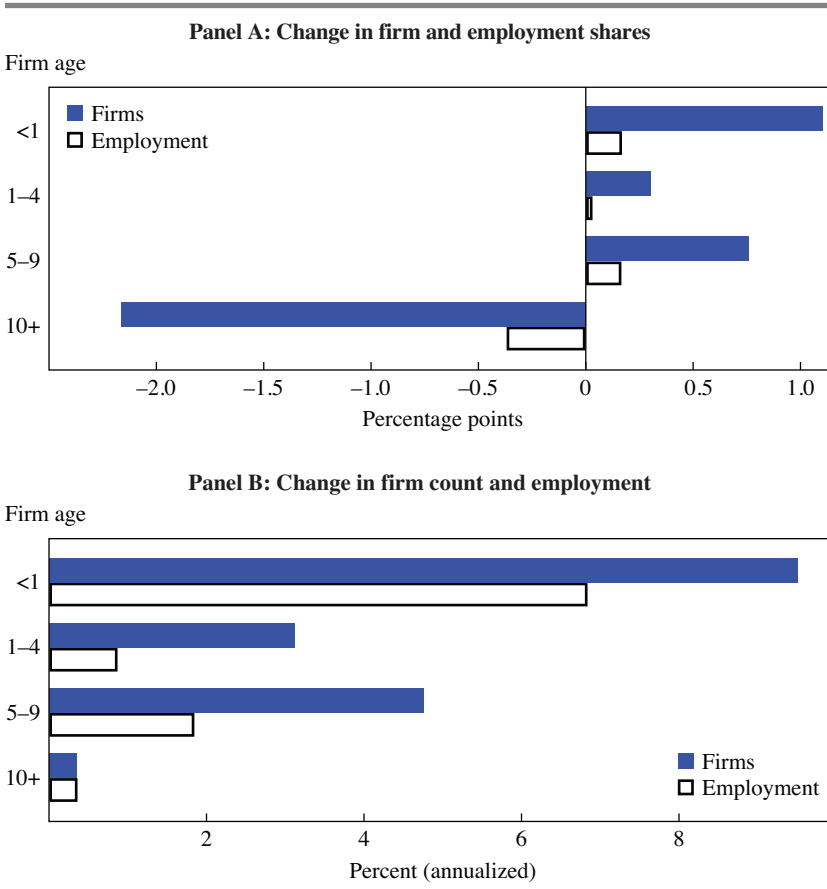
reallocation, where cells are categories that can be defined in terms of firm age groups, firm size groups, geographic divisions, or industries; details are provided in online appendix D, but we provide an overview here. We find that between-cell excess job reallocation increased substantially in the pandemic, especially for cells defined in terms of firm age or firm size by themselves as well as when interacted with spatial or sectoral cells. In other words, we observe a substantial rise in the flow of jobs across these cell boundaries, which implies genuine job reallocation across businesses. The dominant role of reallocation across firm age and firm size groups leads us to explore changes in the firm age and size distribution in the next section.

V.B. Changes in the Firm Age and Size Distribution

The evidence on reallocation—and especially between-cell excess reallocation—implies an increase in the reallocation of activity across businesses in the pandemic. While the changes in the magnitudes of between-cell excess reallocation are large in percentage terms, they are relatively small in terms of absolute flows of jobs. We know from Decker and others (2016), Decker, Haltiwanger, and others (2020), and Karahan, Pugsley, and Şahin (2019) that an important source of the decline in indicators of business dynamism is the shift in activity toward large, mature firms: young and small firms are inherently more dynamic, so the decline in the share of the economy accounted for by young and small firms underlies a significant fraction (albeit far from all) of the decline in the pace of reallocation. In this context, it is instructive to explore changes in the age and size distribution of activity that occurred in the pandemic; we use annual BED data on activity by firm age and size through March 2022.

Figure 19 reports the change in the firm age distribution from March 2020—the very beginning of the pandemic—through March 2022. Panel A shows the percentage point change in the share of firms (solid bars) and employment (hollow bars) accounted for by each firm age group. Young firms' share of activity has risen a bit during the pandemic (after decades of trend decline); the shift in the share of firms is greater than the shift in employment, which is not surprising since pandemic entrants have been smaller than before the pandemic and because the effect of the surge of business entry on employment shares will inherently take time depending on survival rates and post-entry growth patterns of the new firms. The surge in entry has clearly left a mark on the firm age distribution, but even the share of firms five to nine years old increased; these are not pandemic births but are instead relatively young firms that were born before the pandemic. While the activity share changes in panel A of figure 19 must sum

Figure 19. Changing Firm Age Distribution, March 2020 to March 2022



Source: BLS Business Employment Dynamics (BED).
Note: Firms and firm age defined by EIN.

to zero, panel B of the figure shows the percentage growth in the number of firms (solid bars) and employment over this period; for the 2020–2022 period as a whole, all firm age groups saw absolute growth, but the rate of increase was much higher for younger firms (though the growth rates are not quite monotonic). Again, even the oldest young firm category—those age five to nine years—saw rapid growth, with 5 percent more firms and 2 percent more employment than at the beginning of the pandemic (do not forget, though, that firms naturally progress through the age distribution via the process of aging).

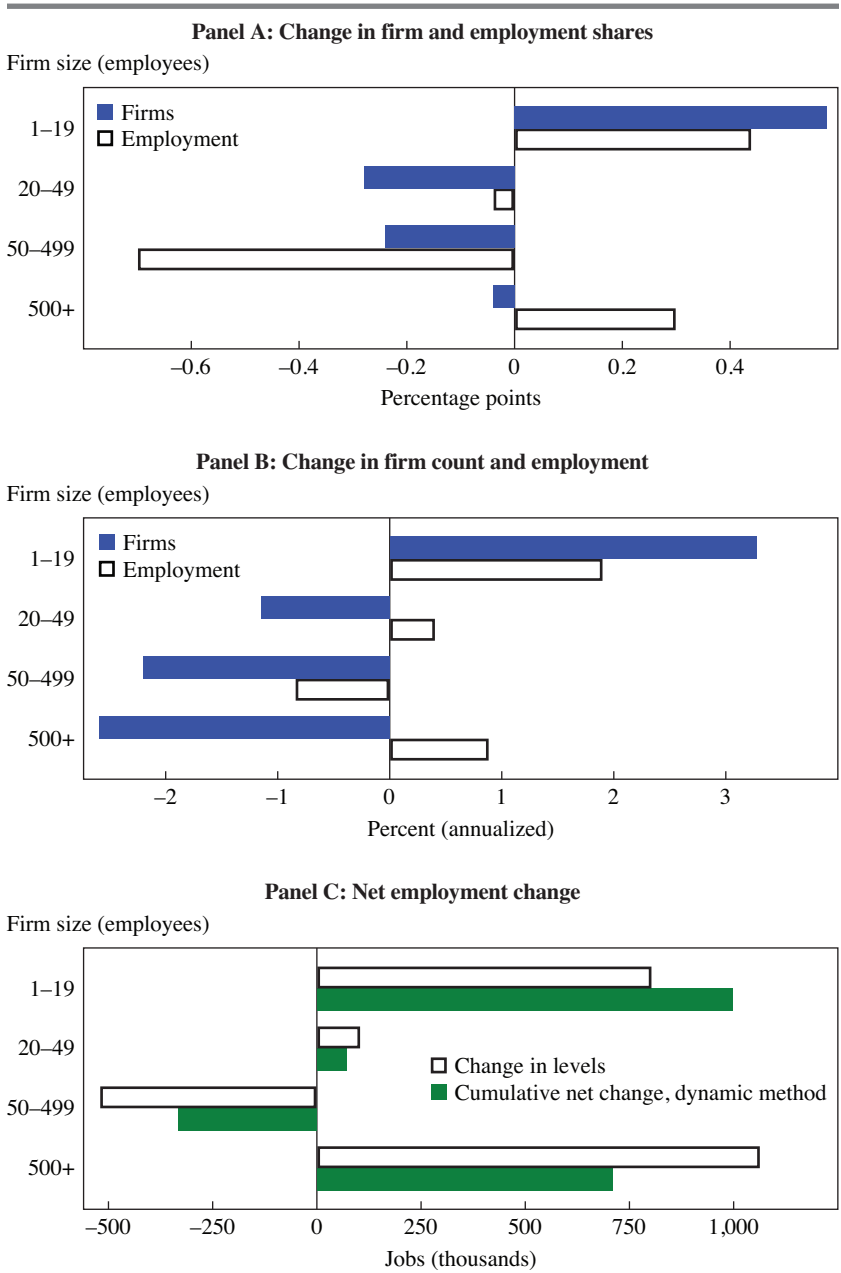
We also examine changes in the firm size distribution. This is more challenging since firms can move both directions through the size distribution; firms with net job destruction may move into smaller size bins, while growing firms may move into larger bins. With this caution in mind, figure 20 reports changes in the size distribution in a manner analogous to figure 19. Panel A shows a shift in the share of firms and employment accounted for by small firms with fewer than 20 employees; but this shift has not been monotonic—firms with between 50 and 499 workers have seen large declines in their share of employment and, especially, firms. In contrast, firms with at least 500 employees have exhibited a modest decline in their share of firms—possibly reflecting firm exit but more likely reflecting firms downsizing into lower bins—but actually saw an increase in their share of employment, as some large firms likely benefited from the pandemic. Panel B, which reports growth in the level of firms and employment, tells a somewhat similar story, with all but the smallest size class seeing a decline in the number of firms but with the largest size class adding jobs. It is important to note that the 1 percent employment growth rate among large firms is substantial given that these firms account for roughly half of all employment, compared with the smallest size class whose share of employment is closer to one-sixth; at the same time, the smallest size class accounts for roughly 90 percent of all firms, so its 3 percent firm count growth rate reflects a large gain in the number of small firms.

As just noted, a challenge associated with firm size distribution analysis is that firms may move either direction across the distribution. But an attractive feature of the BED is that statistics on what BLS denotes as "dynamic sizing" are provided. Dynamic sizing assigns firm job growth to the size bin in which it occurred. For example, if a firm increases from zero employees (i.e., is a firm birth) to thirty-five over a window of time, the first nineteen jobs added are attributed to the 1–19 size class, and the increase from twenty to thirty-five jobs is attributed to the 20–49 size class. Thus, dynamic sizing provides insights into how much of the change in employment observed by size class is due to firms moving across size classes relative to changes within size classes. The BED provides dynamic sizing–based job growth by firm size bin on a quarterly basis.[34]

Panel C of figure 20 reports both the actual change in the level of employment associated with each size bin (hollow bars), which is based on comparing employment levels in March 2022 and March 2020, and

34. See Helfand, Sadeghi, and Talan (2007) for discussion of the BLS dynamic sizing methodology.

Figure 20. Changing Firm Size Distribution, March 2020 to March 2022



Source: BLS Business Employment Dynamics (BED).

Note: Firms defined by EIN. Dynamic method distributes net growth across size categories in which it occurs.

the cumulative dynamic sizing–based employment change (solid bars, constructed by summing quarterly dynamic job flows, by size class, from March 2020 through March 2022). Consider the smallest size class: since the solid bar (dynamic change) is larger than the hollow bar (change in levels), we can infer that there was net movement of firms up and out of this size bin; job growth of firms that graduate out of the size class is (partly) attributed to that size class under dynamic sizing (solid bar) but is not attributed to that class when we simply measure the change in static employment levels (hollow bar). This result for the smallest class is consistent with the surge in firm births, which are typically small, and suggests that some of these firm births—and perhaps also some preexisting small firms—grew out of this size bin. In contrast, for the largest size class, the hollow bar is larger than the solid bar, from which we can infer that there was net movement of firms downward out of this size bin; this is consistent with the net decline in the number of firms in this size class shown in panel B.

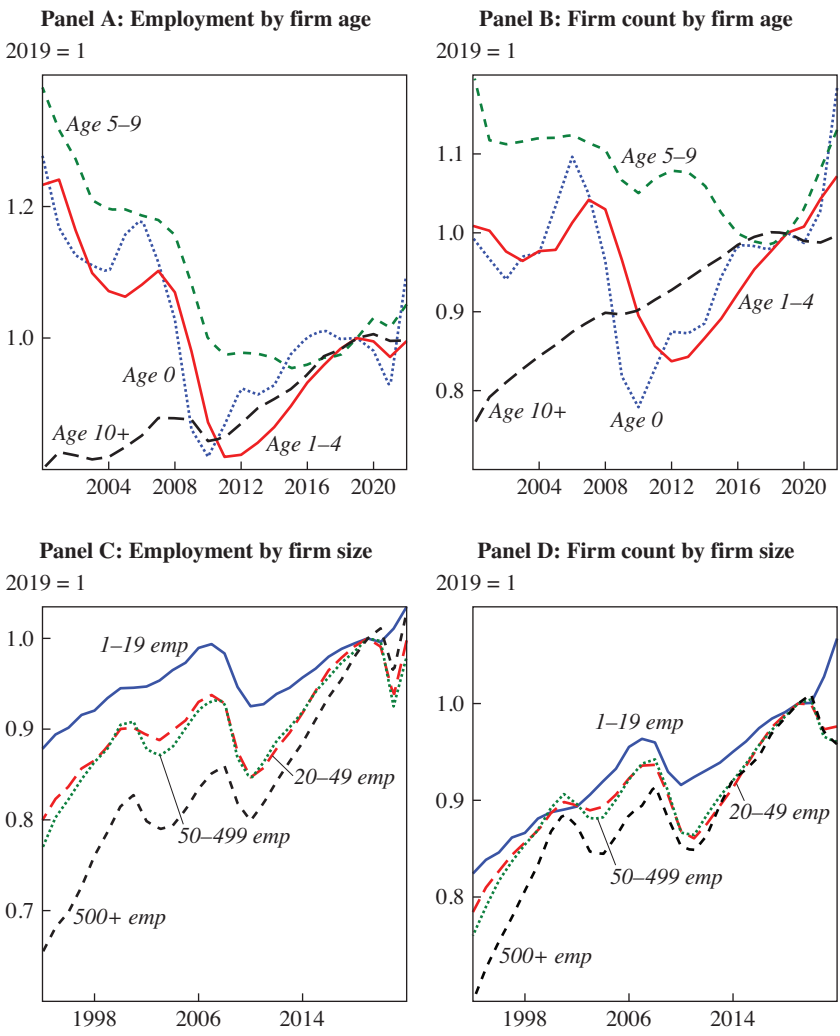
Additional perspective on firm size can be gained by studying the average size of new firm entrants; online appendix figure E14 shows that the average size of new firm entrants in BED data stepped down in the pandemic, consistent with our earlier discussion about unit counts versus employment from entrants. But figure E14 also shows that average entrant size relative to average incumbent size has remained on its pre-pandemic trend; that is, the drop in average entrant size is similar—relative to trend—to the change in average incumbent size. In other words, the relative small size of entrants in the pandemic is not unique to entrants.[35]

These shifts in the firm age and size distribution are remarkable, particularly in a recessionary environment; small and young firms—including young firms born before the pandemic and its dramatic business formation surge—appear to have fared remarkably well during the pandemic. But how much have these shifts reversed the pre-pandemic trends toward mature and large firms? The answer is "not much." Figure 21 depicts the evolution of indexes of employment (panels A and C) and firm counts (panels B and D) by firm age (panels A and B) and firm size (panels C and D). Focusing on panels A and B reporting firm age data, the pre-pandemic shift in activity toward mature firms is evident as the indexes of mature-firm employment (panel A) and firm counts (panel B) rise dramatically during 2000–2020.[36]

35. Online appendix figure E14 also shows BED average size patterns relative to BDS data in the pre-pandemic period; we discuss this in the data appendix.

36. We have confirmed with BLS staffers that there is not a left-truncation bias in the firm age files starting in 2000, despite public-use BED data only starting in 1992, as the BLS has internal micro data affording full accounting of the age 10+ category starting in 2000.

Figure 21. Evolution of Firms and Employment by Age and Size



Source: Business Employment Dynamics (BED).

Note: March snapshots. For age classes above zero, employment measured as implied quarterly DHS denominator. Series indexed to their 2019 value.

In contrast, consistent with the decline in employer business entry, the indexes for young firms (especially firm births at age zero) decline, on net, over this twenty-year period. In the pandemic, these trends begin to reverse—but the decline for mature firms is very modest.

Turning to the evolution of the size distribution (panels C and D of figure 21), the activity shift toward larger firms in recent decades is evident.[37] Again, in the pandemic we have seen some reversal of earlier trends—especially for small firm counts, but not so much on the employment side. Large firms have more employment in 2022 than in 2019, consistent with our evidence above.

Our reading of the data is that there is potentially a beginning of a reversal of the shift in activity to large and mature firms; this is noteworthy and suggests young and small firms weathered the pandemic reasonably well (and, of course, entry has been remarkable). But so far the reversal relative to previous trends is quite modest. A related way to see that the impact has been modest is to compare firm- and employment-weighted entry rates, which we do in online appendix figure E1; there is a notable increase in the firm-weighted start-up rate, but the increase in the employment-weighted start-up rate is less noteworthy.

It is too early to declare an end to the multi-decade decline in business dynamism; such an end will require a sustained increase in employer business entry with, in turn, robust post-entry dynamics (i.e., not a decline in survival rates and post-entry growth conditional on survival). A onetime increase of entry and job reallocation—even if spanning a few years—is different from a persistent elevation of dynamism flows. Still, the striking rise in young and small firm activity in the pandemic is noteworthy.

VI. Taking Stock

Using several official data sources, we document close relationships between business applications, business entry, and job and worker flows during the pandemic. Our findings indicate that the surprising surge in business applications and registrations seen during the pandemic represented genuine entrepreneurial activity and resulted in considerable job creation and reallocation of jobs and workers. This surge in employer entrepreneurship is remarkable given the weakness in broader economic conditions from which it emerged, and it stands in sharp contrast with the plunge in

37. For firm size, we are able to start in 1994 rather than 2000, given we do not face left truncation of the firm size measure as we do for the firm age measure.

employer entrepreneurship seen during the Great Recession. The increase in entrepreneurial activity left its mark on the firm age and size distributions, with a higher share of activity accounted for by young and small firms.

Our findings are consistent with the surging applications yielding increasing new employer businesses. However, it is still too early to study these transitions directly, a task that will require micro data not currently available: the micro data will permit studying applications that transitioned into employer start-ups with a focus on characteristics like industry, location, and entrepreneur demographics, along with post-entry life cycle dynamics. Investigating the demographic patterns of pandemic entrepreneurship looks to be of considerable interest; for example, Fazio and others (2021) find that at the zip code level African American population is strongly predictive of business registrations, so the pandemic may have provided entrepreneurial opportunities to minority groups that have historically faced challenges to business entry.[38]

A related issue that warrants further attention is the high-frequency dynamics of applications and business entry over the course of the pandemic. As we have noted, the surge in applications came in two waves: an initial short-lived wave in the summer (especially July) of 2020, then a second, still ongoing wave commencing in early 2021. It may be that these two waves reflect different incentives and dynamics. The first wave may reflect the distinct market opportunities that arose just after the onset of the pandemic (e.g., online retail), but it may also reflect an increase in nascent entrepreneurship or entrepreneurial brainstorming. Many people found themselves with extra free time in the summer of 2020, given avoidance of high-contact leisure activities and time savings from fewer commutes; some may have used that time—along with broader reassessment of career goals—to consider starting a business. In some cases, these early entrepreneurial ideas may have been overtaken by the (partial) return to more normal patterns of work and leisure later in 2020, and, indeed, we find that the BFS projected firm birth series jumps less than simpler application count series and features a smaller surge in the retail sector. In contrast, in 2021, vaccines started becoming available and pathways out of pandemic isolation were becoming increasingly clear as the country gradually

38. In pre-pandemic data, Dinlersoz and others (2023) find that census tracts with higher African American shares of population have higher application rates but lower transition rates to becoming employers. The latter effect dominates so that census tracts with higher African American shares of the population have lower employer start-up rates per capita. It is of great interest to know whether these distinct patterns of applications and transitions changed in the pandemic.

transitioned toward a post-pandemic new normal. Potential entrepreneurs had more information to plan and start serious businesses by 2021 and this has continued through 2023. We raise these issues since it may be that the transition dynamics of applications to new businesses are very different across these waves. We still lack the data to rigorously discern this distinction, but we do find preliminary evidence for lower transition rates in the early wave, which we discuss in online appendix B.

Our strongest evidence on the surge in business entry is from data on gross and net establishment entry, which includes both new firms and new establishments (new operating locations) of incumbent firms. We find a large and sustained increase in aggregate gross and net establishment entry through 2023:Q1, and the industries and locations with the largest increases in gross and net establishment entry tend to have the largest increases in new business applications. Our evidence on firm entry is consistent with these patterns but is only available through 2022:Q1 and with less industry and spatial detail.

The incentives for new business opportunities induced by the pandemic and its aftermath apply to both new and existing firms, but is the distinction important? Both types of establishment entry are inherent components of reallocation of business activity across the economy, but historically, rapid post-entry growth and innovation are more associated with new firms than with new establishments of existing firms.[39]

Our findings also raise questions about the role of pandemic policies that strongly supported aggregate demand and eased credit conditions—which may be expected to boost firm entry—while also subsidizing incumbent firms via the PPP, the Main Street Lending Program, and the Federal Reserve's corporate credit facilities; Decker, Kurtzman, and others (2020) find that these business support policies included virtually the entire (incumbent) business distribution in their nominal scope for firm size, industry, and legal form. We must leave these and related questions

39. We have some preliminary evidence that this distinction is important. The BED annual files that currently run through 2022:Q1 permit computing establishment entry for establishments less than one year old and for firms less than one year old (where by construction the establishments are also less than one year old). From 2019:Q1 through 2022:Q1, annual total establishment births (i.e., age less than one year) rose by 38 percent, while the annual number of establishments of new firms grew at 21 percent (the latter is consistent with the firm entry rates reported in online appendix figure E1). Both are substantial, but the higher growth of total establishment births suggests an important role for new establishments at incumbent firms. Notably, though, we also find total establishment births grew more rapidly from 2020:Q1 to 2021:Q1 than establishments at firm births, suggesting establishment entry for existing firms was more resilient early in the pandemic than firm births.

for future research, which we hope will be informed by the large collection of facts we have assembled. In the meantime, our existing results suggest that entrepreneurship has played a key role in pandemic-era labor market dynamics.

One topic that is conspicuously missing from our analysis is an investigation of the surge in business applications that are likely nonemployers. Per Bayard and others (2018), likely nonemployer applications have a very low probability of becoming employer businesses (about 3 percent), and prior to the pandemic these applications tracked nonemployer activity reasonably well (Haltiwanger 2022).[40] Given the very large increase in likely nonemployer applications, the increase in entrepreneurship may be substantially greater than we have characterized via the potential increase in new nonemployer businesses. But the Nonemployer Statistics (NES) from the Census Bureau are currently available only through 2020. An alternative path is to use the Current Population Survey (CPS) or other household surveys that track self-employment activity; but there has been a growing discrepancy between self-employment activity tracked by the administrative data, such as the NES, and household data (Abraham and others 2021). Relatedly, the nonemployers of relevance to the BFS are those with an EIN, but most nonemployers do not have an EIN. Nonemployers with EINs are substantially larger than those without an EIN; only 15 percent of sole proprietors have EINs, and the small sole proprietors without EINs are dominated by individuals for whom nonemployer activity is supplemental (often to a wage and salary income) or reflects stopgap activity.[41] Published NES data do not separately tabulate sole proprietors with and without EINs, and the CPS only distinguishes between incorporated and non-incorporated self-employed. In short, there are challenges to investigating the implied dynamics of the surge in likely nonemployers. But given the magnitude of the increase in likely nonemployer applications (see figure 1), exploring this topic is of considerable interest; moreover, there has been much discussion of the pandemic changing attitudes toward work, including the recognition that important tasks can be done remotely. And an argument could be made that the nonpecuniary benefits of being one's own boss—as discussed in Hurst and Pugsley (2011)—may have risen. A potential implication is that individuals have increasingly decided to go out on their own as nonemployers, but at this point nonemployer measurement is limited.

40. See also online appendix A.
41. See Davis and others (2009) and Abraham and others (2021).

VII. Implications for the Future?

Given that we are only beginning to observe the real activity effects connected to the surge in new business applications, discussion of the implications of this surge for the future of US economic activity can only be highly speculative. Thus, here we provide some discussion about what potential patterns are worth contemplating in the coming months and years.

First, we emphasize that the full implications of the pandemic start-up surge will take several years to unfold. This reflects the highly volatile nature of start-ups, especially over their first five to ten years. Most start-ups fail or, at least, do not grow (Decker and others 2014). A small fraction grow rapidly, and this small subset of entrants is disproportionately important for the contribution of start-ups to job creation, innovation, and productivity growth (Decker and others 2014; Guzman and Stern 2020; Sterk, Sedláček, and Pugsley 2021). Theory and evidence suggest that start-ups are a core part of the experimentation that accompanies the development and adoption of new technologies and production processes, though this experimentation necessarily involves many business failures (Foster and others 2021).

Second, this increase in start-ups has occurred in spite of factors that were dampening the pace of business entry—and business dynamism more generally—in the decades leading up to the pandemic (Decker, Haltiwanger, and others 2020). It is unlikely that those factors, while still not completely understood, have disappeared entirely. Whether the countervailing forces driving the pandemic surge are sufficient to change the pre-pandemic trend decline is unclear; as we discuss in section V, the shock to entry and reallocation seen during the pandemic would have to be very persistent, and the new cohorts of entrants would have to feature a sufficient number of high-growth firms, for past trends to be substantially reversed.

Third, it may be important to consider the dynamics of aggregate productivity prior to the pandemic. In online appendix figures E21 and E22, the well-known productivity slowdown in the post-2005 period, and especially since 2014, is evident even in the innovative high-tech sectors of the economy. Many factors have been proposed as underlying this slowdown—including the decline in dynamism and entrepreneurship (Decker, Haltiwanger, and others 2020)—so the pandemic-era pattern of business formation may have implications for how productivity evolves going forward.

This discussion suggests some possible implications of the pandemic business entry surge. One possibility is that this surge is associated with

a burst of innovation, with start-ups being an important component of the experimentation leading to that innovation. Hints of this possibility may be seen in the industry composition of surging applications and establishment openings (online appendix figure E10), with high-tech industries like nonstore retail, software publishing, computer systems design, scientific research and development services (e.g., AI businesses), and data processing apparently seeing especially elevated entry. While the evidence on actual new employer businesses in high-tech industries is still emerging, high-tech industries have the highest pace of projected start-ups of any broad sector through September 2023. Tracking the potential for surging entrepreneurship to spark economic growth and technological progress should be a high priority; eventually we would hope to see such progress reflected in productivity statistics, and a productivity boost from surging start-ups could mean stronger growth of potential output for the economy overall. Again, it will take some time for these dynamics to unfold, but early signals of the nature and composition of this surge might be detected, for example, using the nowcasting methodology of Guzman and Stern (2017).

Alternatively, this surge may reflect the type of spatial and sectoral restructuring that we have detected—but only insofar as such restructuring is necessary for providing basic support activities for the changing nature of work and lifestyle, with no broader spillovers in terms of innovation, productivity, and growth. In other words, the surge in start-ups suggested by the data we have reviewed could reflect a reshuffling of economic activity without leading to additional technological progress or growth. The surge of entrants in the service industries (e.g., restaurants and gyms) is consistent with this perspective. And the within-city donut effects we (and others) observe in the spatial patterns of applications and actual increases in net establishment growth may reflect business formation to support the increased fraction of working hours spent at home, and little else. Such support activity is likely very important to enable the changing nature of work—to the extent that the change is persistent—but it is unclear that such reallocation would herald a burst of innovation and productivity growth. A related possibility is that the pandemic presented a shock to entrepreneurial preferences, as in Hurst and Pugsley (2011); this is consistent with the drop in average entrant size. Whether persistent or not, such a shock is also unlikely to be associated with a burst of innovation and productivity growth.

Finally, we acknowledge the widely speculated upon possibility of an economic slowdown. Since early 2022, US monetary policy has tightened

materially in response to elevated inflation, and financial condition measures are now much more restrictive than they were in the early pandemic period (Ajello and others 2023). While business applications have remained reasonably stable at their elevated pandemic level through September 2023 (see figure 1), monetary policy is typically thought to operate with long and variable lags. Existing literature—for example, Davis and Haltiwanger (2021)—finds that start-ups and young businesses are particularly sensitive to business cycle fluctuations, particularly those associated with tight financial conditions (e.g., falling house prices, rising interest rates, or declining business lending activity). The young businesses started during the pandemic, and the continued elevated trend of business applications, may be at risk in the event of a broad economic slowdown.

ACKNOWLEDGMENTS The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors of the Federal Reserve System. We thank Janice Eberly, Jorge Guzman, Ben Pugsley, Scott Stern, participants at the Annual Research Conference at the Boston Federal Reserve, and participants at the Fall 2023 *BPEA* Conference for comments on earlier drafts of this paper. We thank Aditya Pande and Matilde Serrano for excellent research assistance and the Templeton Foundation for financial support. We thank Eric Simants and Kevin Cooksey for fielding numerous questions about Business Employment Dynamics data and for providing vintage files, though any errors in data use and interpretation are our own. This paper uses public domain data from the Bureau of Labor Statistics and the US Census Bureau.

References

Abraham, Katharine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2021. "Measuring the Gig Economy: Current Knowledge and Open Issues." In *Measuring and Accounting for Innovation in the Twenty-First Century*, edited by Carol Corrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel. Chicago: University of Chicago Press.

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr. 2018. "Innovation, Reallocation, and Growth." *American Economic Review* 108, no. 11: 3450–91.

Ajello, Andrea, Michele Cavallo, Giovanni Favara, William B. Peterman, John W. Schindler IV, and Nitish R. Sinha. 2023. "A New Index to Measure U.S. Financial Conditions." FEDS Notes. Washington: Board of Governors of the Federal Reserve System.

Akcigit, Ufuk, and Nathan Goldschlag. 2023. "Where Have All the 'Creative Talents' Gone? Employment Dynamics of US Inventors." Working Paper 31085. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31085.

Akcigit, Ufuk, and William R. Kerr. 2018. "Growth through Heterogeneous Innovations." *Journal of Political Economy* 126, no. 4: 1374–443.

Alon, Titan, David Berger, Robert Dent, and Benjamin Wild Pugsley. 2018. "Older and Slower: The Startup Deficit's Lasting Effects on Aggregate Productivity Growth." *Journal of Monetary Economics* 93:68–85.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *Quarterly Journal of Economics* 135, no. 2: 645–709.

Bayard, Kimberly, Emin Dinlersoz, Timothy Dunne, John C. Haltiwanger, Javier Miranda, and John Stevens. 2018. "Early-Stage Business Formation: An Analysis of Applications for Employer Identification Numbers." Working Paper 24364. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w24364.

Breaux, Cory, and Alisha Gurnani. 2022. "PPP-BFS Project," presentation, US Bureau of the Census, September.

Cajner, Tomaz, Leland D. Crane, Ryan A. Decker, John Grigsby, Adrian Hamins-Puertolas, Erik Hurst, Christopher Kurz, and Ahu Yildirmaz. 2020. "The US Labor Market during the Beginning of the Pandemic Recession." *Brookings Papers on Economic Activity*, Summer, 3–33.

Choi, Joonkyu, Samuel Messer, Michael A. Navarrete, and Veronika Penciakova. 2023. "Unemployment Benefits Expansion and Business Formation." Working Paper. https://www.dropbox.com/scl/fi/5ikquvssbwz9nn00yt4bo/UI_BFS.pdf?rlkey=f3wm9feuedlgpe5b4b3zgs8cz&dl=0.

Crane, Leland D., Ryan A. Decker, Aaron Flaaen, Adrian Hamins-Puertolas, and Christopher Kurz. 2022. "Business Exit during the COVID-19 Pandemic: Non-traditional Measures in Historical Context." *Journal of Macroeconomics* 72: 103419.

Davis, Steven J., R. Jason Faberman, and John C. Haltiwanger. 2012. "Labor Market Flows in the Cross Section and Over Time." *Journal of Monetary Economics* 59, no. 1: 1–18.

Davis, Steven J., and John C. Haltiwanger. 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." *Quarterly Journal of Economics* 107, no. 3: 819–63.

Davis, Steven J., and John C. Haltiwanger. 2014. "Labor Market Fluidity and Economic Performance." In *Economic Policy Symposium Proceedings: Re-evaluating Labor Market Dynamics*. Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Davis, Steven J., and John C. Haltiwanger. 2021. "Dynamism Diminished: The Role of Housing Markets and Credit Conditions." Working Paper 25466. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w25466.

Davis, Steven J., John C. Haltiwanger, Ronald S. Jarmin, C. J. Krizan, Javier Miranda, Alfred Nucci, and Kristin Sandusky. 2009. "Measuring the Dynamics of Young and Small Businesses: Integrating the Employer and Nonemployer Universes." In *Producer Dynamics: New Evidence from Micro Data*, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. Chicago: University of Chicago Press.

Davis, Steven J., John C. Haltiwanger, and Scott Schuh. 1996. *Job Creation and Destruction*. Cambridge, Mass.: MIT Press.

Decker, Ryan A., and John C. Haltiwanger. 2022. "Business Entry and Exit in the COVID-19 Pandemic: A Preliminary Look at Official Data." FEDS Notes. Washington: Board of Governors of the Federal Reserve System.

Decker, Ryan A., John C. Haltiwanger, Ronald S. Jarmin, and Javier Miranda. 2014. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives* 28, no. 3: 3–24.

Decker, Ryan A., John C. Haltiwanger, Ronald S. Jarmin, and Javier Miranda. 2016. "Where Has All the Skewness Gone? The Decline in High-Growth (Young) Firms in the U.S." *European Economic Review* 86:4–23.

Decker, Ryan A., John C. Haltiwanger, Ronald S. Jarmin, and Javier Miranda. 2020. "Changing Business Dynamism and Productivity: Shocks versus Responsiveness." *American Economic Review* 110, no. 12: 3952–90.

Decker, Ryan A., Robert J. Kurtzman, Byron F. Lutz, and Christopher J. Nekarda. 2020. "Across the Universe: Policy Support for Employment and Revenue in the Pandemic Recession." Finance and Economics Discussion Series. Washington: Board of Governors of the Federal Reserve System.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." *Quarterly Journal of Economics* 135, no. 2: 561–644.

Dinlersoz, Emin, Timothy Dunne, John C. Haltiwanger, and Veronika Penciakova. 2021. "Business Formation: A Tale of Two Recessions." *American Economic Association: Papers and Proceedings* 111:253–57.

Dinlersoz, Emin, Timothy Dunne, John C. Haltiwanger, and Veronika Penciakova. 2023. "The Local Origins of Business Formation." Working Paper CES-23-34. Washington: Center for Economic Studies.

Duguid, James, Bryan Kim, Lindsay Relihan, and Chris Wheat. 2023. "The Impact of Work-from-Home on Brick-and-Mortar Retail Establishments: Evidence from Card Transactions." Working Paper. Social Science Research Network, June 5. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4466607.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2010. "The Labor Market in the Great Recession." *Brookings Papers on Economic Activity*, Fall, 1–48.

Fairlie, Robert. 2020. "The Impact of COVID-19 on Small Business Owners: Evidence from the First Three Months after Widespread Social-Distancing Restrictions." *Journal of Economics and Management Strategy* 29, no. 4: 727–40.

Fazio, Catherine E., Jorge Guzman, Yupeng Liu, and Scott Stern. 2021. "How Is COVID Changing the Geography of Entrepreneurship? Evidence from the Startup Cartography Project." Working Paper 28787. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w28787.

Fazio, Catherine E., Jorge Guzman, and Scott Stern. 2020. "New Business Needs a Big Rescue, Too." *Forbes*, April 24. https://www.forbes.com/sites/columbiabusinessschool/2020/04/24/new-business-needs-a-big-rescue-too/?sh=12da5735309c.

Federal Reserve System Board of Governors. 2020. *Monetary Policy Report June 2020*. Washington: Author.

Foster, Lucia, Cheryl Grim, John C. Haltiwanger, and Zoltan Wolf. 2021. "Innovation, Productivity Dispersion, and Productivity Growth." In *Measuring and Accounting for Innovation in the Twenty-First Century*, edited by Carol Carrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel. Chicago: University of Chicago Press.

Guzman, Jorge, and Scott Stern. 2017. "Nowcasting and Placecasting Entrepreneurial Quality and Performance." In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, edited by John C. Haltiwanger, Erik Hurst, Javier Miranda, and Antoinette Schoar. Chicago: University of Chicago Press.

Guzman, Jorge, and Scott Stern. 2020. "The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 32 US States, 1988–2014." *American Economic Journal: Economic Policy* 12, no. 4: 212–43.

Haltiwanger, John C. 2016. Comment on "Understanding Declining Fluidity in the U.S. Labor Market," by Molloy, Raven, Christopher L. Smith, Riccardo Trezzi, and Abigail Wozniak. *Brookings Papers on Economic Activity*, Spring, 241–52.

Haltiwanger, John C. 2020. "Applications for New Businesses Contract Sharply in Recent Weeks: A First Look at the Weekly Business Formation Statistics." Unpublished.

Haltiwanger, John C. 2022. "Entrepreneurship during the COVID-19 Pandemic: Evidence from the Business Formation Statistics." *Entrepreneurship and Innovation Policy and the Economy* 1:9–42.

Haltiwanger, John C., Henry R. Hyatt, Lisa B. Kahn, and Erika McEntarfer. 2018. "Cyclical Job Ladders by Firm Size and Firm Wage." *American Economic Journal: Macroeconomics* 10, no. 2: 52–85.

Hansen, Stephen, Peter John Lambert, Nicholas Bloom, Steven J. Davis, Raffaella Sadun, and Bledi Taska. 2023. "Remote Work across Jobs, Companies, and Space." Working Paper 31007. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31007.

Helfand, Jessica, Akbar Sadeghi, and David Talan. 2007. "Employment Dynamics: Small and Large Firms over the Business Cycle." *Monthly Labor Review*, March, 39–50.

Hurst, Erik, and Benjamin Wild Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity*, Fall, 73–118.

Karahan, Fatih, Benjamin Wild Pugsley, and Ayşegül Şahin. 2019. "Demographic Origins of the Startup Deficit." Working Paper 25874. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w25874.

Molloy, Raven, Christopher L. Smith, Riccardo Trezzi, and Abigail Wozniak. 2016. "Understanding Declining Fluidity in the U.S. Labor Market." *Brookings Papers on Economic Activity*, Spring, 183–237.

O'Brien, Connor. 2022. "More Physical Places of Businesses Open Now than Pre-Pandemic, Led by Sun Belt Metros." Economic Innovation Group, April 5. https://eig.org/more-physical-places-of-businesses-open-now-than-pre-pandemic-led-by-sun-belt-metros/.

Ramani, Arjun, and Nicholas Bloom. 2021. "The Donut Effect: How COVID-19 Shapes Real Estate." Policy Brief. Stanford, Calif.: Stanford Institute for Economic Policy Research.

Rosenberg, Eli. 2022. "4.3 Million Americans Left Their Jobs in December as Omicron Variant Disrupted Everything." *Washington Post*, February 1. https://www.washingtonpost.com/business/2022/02/01/job-quits-resignations-december-2021/.

Sterk, Vincent, Petr Sedláček, and Benjamin Wild Pugsley. 2021. "The Nature of Firm Growth." *American Economic Review* 111, no. 2: 547–79.

Comment and Discussion

COMMENT BY

JORGE GUZMAN Ryan Decker and John Haltiwanger bring to this issue of *BPEA* a thought-provoking piece on the evolution of US entrepreneurship after the COVID-19 pandemic. Using multiple US Census Bureau data sets they present systematic evidence that the level of firm formation for both employer and nonemployer firms increased after COVID-19. This increase is large and, at least up to the time of writing, persistent. Importantly, the rise in entrepreneurship comes as a much needed respite to the long drop in the quantity of young firms previously documented by the authors (Decker and others 2014). At least within their census data, it is the first substantial increase in the number of new firms since 1977, the earliest year available. Other data sets unrelated to the census have also documented an increase in entrepreneurship after COVID-19, most notably business registration statistics using state-level registries (Fazio and others 2021), suggesting that the increase documented by Decker and Haltiwanger is real.

The bulk of my discussion focuses on two questions. First, what is causing this boom in new firm formation? Second, what does such a large increase in new firms imply for the economy? Neither question has a clear answer, but the gap is particularly salient for the latter one. The inability to answer these questions emphasizes how nascent our understanding of the role of entrepreneurship in the economy is, making it fertile ground for future research.

Editors' Note: Benjamin Pugsley provided a thoughtful discussion on the conference version of the paper by Decker and Haltiwanger at the Fall 2023 *BPEA* Conference. The recording of his discussion can be found at https://www.brookings.edu/events/bpea-fall-2023-conference/.

Brookings Papers on Economic Activity, Fall 2023: 303–316

WHY DID ENTREPRENEURSHIP INCREASE AFTER COVID-19?

Is entrepreneurship rising due to higher business dynamism and creative destruction? For economists, the most valuable benefit of entrepreneurship to the economy is its crucial role in productivity growth. This occurs through two channels: business dynamism, or the reallocation of labor and capital from less productive to more productive firms even within narrowly similar product categories (Decker and others 2014); and creative destruction, or the process through which the desire for profits leads to process, product, and organizational inventions that incorporate a de novo way of doing economic activities (Schumpeter 1943; Akcigit and Kerr 2018; Acemoglu and Robinson 2013). The line between these two activities is not clearly defined. Many cases may imply both, and some economists have used the terms *business dynamism* and *creative destruction* interchangeably. However, they refer to different sources of variation on the nature of productivity growth. Business dynamism more closely relates to the efficient allocation of capital and labor across existing projects in the economy, while creative destruction, even if resulting in business dynamism, focuses more on the way profit motives promote investment for the development of new technologies, organizations, and business models.

Both within and outside the paper, evidence is consistent with both effects being partly responsible for the changes in US entrepreneurship after COVID-19.

Consider creative destruction first. By looking at changes within individual industries, Decker and Haltiwanger show in figure 3 that changes in industry composition related to technological innovation are taking place. Some of these industry changes are temporary (e.g., the need for more personal and health care services in 2020 or the supply chain struggles of 2021). However, by 2021, we observe what appears to be a partial reorganization of the economy: the founding of new nonstore retailers (e-commerce) has more than doubled, and new firm start-ups in the professional, scientific, and technical category, which includes the majority of those typically called tech firms, has also increased.

Other evidence outside the paper also supports this hypothesis. In particular, there was a boom in venture capital financing during the COVID-19 years, which in 2022 reached its highest levels since the year 2000. Research has documented clearly that venture capital booms lead to the financing and growth of more innovative ideas (Howell and others 2020; Nanda and Rhodes-Kropf 2013), making it possible that the current wave of new

innovations, such as artificial intelligence or commercial space travel, creates a more productive organization of the economy.

Next, consider business dynamism. Beyond innovation incentives, do we observe economic activity reallocating from less productive to more productive firms?

Here, it is useful to remind ourselves of the details of the economic moment in which the boom in entrepreneurship occurred. In the period after COVID-19, employee quit rates increased despite strong economic fundamentals, leading to a phenomenon sometimes called the Great Resignation. By 2022, for example, employee quit rates were 50 percent higher than would have been predicted by models based on economic fundamentals (Gittleman 2022). At the same time, existing firm sales dropped precipitously, by up to 40 percent in 2021 (Barrero and others 2021), while labor force participation appears, if anything, to have increased (Sheiner and Salwati 2022). Put simply, the economy is robust and there are a substantial number of jobs, but incumbent firms are not doing well and workers are leaving them quickly. Where is all this labor to flow? The most likely possibility is new entrants, that is, entrepreneurship.

Decker and Haltiwanger present evidence that appears consistent with this story. In figure 9, for example, they show that establishment exits have been increasing concurrently with entry. In figure 11, they show that excess entry has occurred in virtually all sectors of the economy.

Overall, the evidence suggests an increase in business dynamism and business reallocation. However, it is also fundamental to ask why individuals have increased their preference to become entrepreneurs.

Is there a changing utility value of entrepreneurship, and could there be a role for work-from-home technology? A different family of explanations does not focus so much on macroeconomic concepts such as creative destruction or dynamism but instead uses a choice-based approach to consider why some individuals would leave wage employment for the opportunity to start a new firm. When one considers the typical US resident's utility function, what is entrepreneurship's role in maximizing utility, and has this changed? Explanations considering this argument focus on two separate shocks through COVID-19. First, they emphasize that the COVID-19 shock and lockdowns, by requiring families to remain at home for extended periods (sometimes making significant changes to their space at home or their living situation), increased the importance individuals placed on being at home or independent. This, in turn, led them to start more firms. Second, the argument also tends to have a technological logic

behind it: the advent of work-from-home (WFH) technologies, particularly videoconferencing, enabled many individuals to remain at home and finally do the independent work that is best suited to them.

Under these utility-based explanations, the economic benefits of the rise in entrepreneurship become more nuanced. Even if the choice to start a firm is utility maximizing, it does not lend itself directly to productivity improvements for the economy. While the once-worker-now-entrepreneur is possibly better off (at least based on revealed preferences), the economy may be the same. Indeed, in extreme cases, the additional focus on independence and leisure may lead to a productivity slowdown, in which the economy is composed of too many small firms that do not scale due to utility-driven growth frictions (Hamilton 2000).

For existing workers, the hypothesis that WFH technologies increased entrepreneurship does not appear consistent with research on the impact of information technology (IT). In particular, the presumed role of WFH technologies in enabling a large portion of new home-based businesses seems less likely, because even though WFH technologies certainly increased the possibility of starting a business at home, its most significant impact was in enabling the possibility of working from home as the employee of a company. The main utility benefits of entrepreneurship, such as freedom and time flexibility (Hamilton 2000), being close to home (Rosenthal and Strange 2012), or being away from one's boss, have become relatively much more accessible to company employees. Given evidence that IT typically supports a decentralization of decision authority and an emphasis on subjective incentives, both of which seem complementary to working from home, the most realistic prediction would instead be a reduction in new firm formation and a boom in jobs in big corporations, as a large share of both existing and new workers find a series of jobs (previously inaccessible) that give the freedom they seek.

Yet, this argument is only half the picture. To the extent that worker preferences also changed toward being an entrepreneur by valuing freedom and flexibility more, or that WFH technologies allowed individuals previously out of the labor force to reenter the economy, then the overall incidence of entrepreneurship could increase.

A different potential channel for WFH technologies involves changing the boundary of the firm, allowing some transactions that used to take place in a firm to be done through the market (Forman and McElheran 2019). This is the case, for example, with gig workers on platforms such as Uber and Taskrabbit, both of which created many small-scale entrepreneurs who provide services to the platform or use gig work as a baseline to start firms

on their own (Barrios, Hochberg, and Yi 2022). The possibility that these platforms enabled additional online services is still to be investigated.

Finally, bringing back the possibility of changing worker preferences, there may be individual changes in the types of jobs people are willing to accept. Besides the COVID-19 pandemic, the year 2020 was witness to one of the largest social movements since the civil rights era, Black Lives Matter, leading one to ask whether minority groups might have more directly experienced a change in the way they think through or choose their career.

Is there a rise in entrepreneurship for minorities? Building on the results presented in earlier work by Haltiwanger, in Fazio and others (2021) my coauthors and I use business registration records to document a significant heterogeneity in the changing geography of entrepreneurship after COVID-19.[1] Our key result is that this heterogeneity does not merely reflect the gradual transition of individuals out of central business districts into the suburbs but instead is statistically related to race: zip codes with a high share of Black residents have the highest increases in entrepreneurship. Other variables such as income, population density, or age hold no relationship. The impact is even more striking when one considers contiguous zip codes within a city. For example, in maps of New York City, we can consider changes in entrepreneurship across neighborhoods that are adjacent but have significantly different racial compositions, such as Central Harlem, Morningside Heights, and Washington Heights. There are substantial differences among them, with Harlem clearly having a larger increase in new firms compared to others. This pattern is also apparent in this paper by Decker and Haltiwanger. In their state-level map (figure 4), we observe that the largest increases are in the Deep South, including states typically low on productivity, such as Alabama and Mississippi. These increases surpass other states that saw ample in-migration during COVID-19 by people expecting to work from home, such as Florida, Arizona, Texas, and Tennessee. All of this suggests the possibility that the increase in entrepreneurship after COVID-19 is related to the incidence of Black population across regions.

There are at least three mechanisms for such an increase. The first possibility is a change in local demand. Since the pandemic created a significant movement of people, these new residents would now create local demand in new neighborhoods. Such an explanation does not appear readily consistent with the empirical patterns. While pandemic reallocation happened out of business districts and toward lower-density areas, Black neighborhoods

1. See, for example, Haltiwanger (2022).

are in more dense locations than white neighborhoods. Zip code population density also does not predict the increase in entrepreneurship rates in our analysis.

A second group of explanations instead relates to more behavioral aspects associated with changes in the demand, jobs, or general expectations for potential Black business owners. Bennett and Robinson (2023) document significant differences in business practices across race, which in turn can be influenced by social movements that co-occurred with the COVID-19 pandemic, such as Black Lives Matter. This appears an important question much in need of empirical evidence.

Finally, a third (and clearer) option is that COVID-19 ultimately brought differences in financial access.

Has there been improved financial access for minorities after COVID-19? There are at least two mechanisms through which the COVID-19 pandemic could have increased financial access. One is government intervention; the other is technological change through fintechs. To understand the logic of both it is important to recognize the differences that exist in the incidence of financial institutions across neighborhoods and race. As documented by Small and others (2021), predominantly Black neighborhoods tend to be farther away from conventional retail banks, making traditional access to financing harder. Policies that reduce such geographic inequality can be particularly valuable for new investments, including new firms.

Consider the government interventions during the pandemic: the COVID-19 stimulus package was, to a large extent, equally distributed across neighborhoods, ameliorating disparate access to financing due to geography. In Fazio and others (2021), we also show a measurable increase in entrepreneurship in the few weeks after the American Rescue Plan (Biden stimulus).

In the case of fintechs, the key possibility is that because online banking companies are less locally determined, they may be able to access areas that are not typically well banked. Erel and Liebersohn (2022) show that fintech banks are more likely to serve minority households and locations with fewer bank branches. Chernenko and Scharfstein (2022) find there were wide racial disparities in the Paycheck Protection Program, which are at least partially ameliorated by fintechs.[2] In essence, the transition to more online banking after COVID-19 may have had a positive influence for previously underbanked neighborhoods.

2. See also Griffin, Kruger, and Mahajan (2023).

WHAT IS THE IMPACT OF A BOOM IN ENTREPRENEURSHIP ON THE ECONOMY? Moving beyond the causes of the rise in entrepreneurship after COVID-19 to the consequences, it is only natural to ask what are the economic implications of this massive increase in new firms.

Here, one can't help but be surprised at the level of uncertainty that comes with these predictions. Even though the rise in entrepreneurship during COVID-19 is the largest increase in our lifetimes, the predictions drawn from this increase by the authors and other entrepreneurship economists (myself included) are very cautious. We do not know exactly what it means, and we are not sure whether it implies increases in productivity growth, creative destruction, or social equity.

The fact that we are unable to predict outcomes is a symptom of the incompleteness, and opportunity, of entrepreneurship economic theory. Whereas a macroeconomist knows that productivity numbers of 4 percent, 2 percent, or 1 percent are worlds apart from each other in their implications for the economy, or that inflation at 1 percent versus 5 percent would lead to drastically different paths of investment and business activity, entrepreneurship economists do not yet know what to make of the shifts and flows of new firm formation for the economy or even for our own conclusions. While the mechanisms of entrepreneurship are now somewhat appreciated, the way these come together to have an impact on economic growth is not.

CONCLUSION Decker and Haltiwanger present a paper that, like many good papers, opens more questions than it answers. By going through the careful process of simply describing the evolution of measures of new firm formation within the US Census, they leave the reader with the desire to learn a lot more about both the causes and consequences of entrepreneurship in the economy. Large economic shocks, such as the Great Depression, the stagflation of the 1970s, or the collapse of the Soviet Union, have always provided fertile ground for economists to test their theories and, ex post, develop new substantive ones that can better explain the changing economy. The COVID-19 pandemic is likely to be a similar shock, providing much to study regarding the reorganization of the economy, with entrepreneurship being one of the settings in which this takes place.

REFERENCES FOR THE GUZMAN COMMENT

Acemoglu, Daron, and James A. Robinson. 2013. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown.

Akcigit, Ufuk, and William R. Kerr. 2018. "Growth through Heterogeneous Innovations." *Journal of Political Economy* 126, no. 4: 1374–443.

Barrero, Jose Maria, Nicholas Bloom, Steven J. Davis, and Brent H. Meyer. 2021. "COVID-19 Is a Persistent Reallocation Shock." *American Economic Association Papers and Proceedings* 111:287–91.

Barrios, John M., Yael V. Hochberg, and Hanyi Yi. 2022. "Launching with a Parachute: The Gig Economy and New Business Formation." *Journal of Financial Economics* 144, no. 1: 22–43.

Bennett, Victor Manuel, and David T. Robinson. 2023. "Why Aren't There More Minority Entrepreneurs?" Social Science Research Network, February 21. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4360750.

Chernenko, Sergey, and David S. Scharfstein. 2022. "Racial Disparities in the Paycheck Protection Program." Working Paper 29748. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w29748.

Decker, Ryan A., John C. Haltiwanger, Ronald S. Jarmin, and Javier Miranda. 2014. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives* 28, no. 3: 3–24.

Erel, Isil, and Jack Liebersohn. 2022. "Can FinTech Reduce Disparities in Access to Finance? Evidence from the Paycheck Protection Program." *Journal of Financial Economics* 146, no. 1: 90–118.

Fazio, Catherine E., Jorge Guzman, Yupeng Liu, and Scott Stern. 2021. "How Is COVID Changing the Geography of Entrepreneurship? Evidence from the Startup Cartography Project." Working Paper 28787. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w28787.

Forman, Chris, and Kristina McElheran. 2019. "Production Chain Organization in the Digital Age: I.T. Use and Vertical Integration in U.S. Manufacturing." Social Science Research Network, June 13. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3396116.

Gittleman, Maury. 2022. "The 'Great Resignation' in Perspective." *Monthly Labor Review*, July. https://www.bls.gov/opub/mlr/2022/article/the-great-resignation-in-perspective.htm.

Griffin, John M., Samuel Kruger, and Prateek Mahajan. 2023. "Did FinTech Lenders Facilitate PPP Fraud?" *Journal of Finance* 78, no. 3: 1777–827.

Haltiwanger, John C. 2022. "Entrepreneurship during the COVID-19 Pandemic: Evidence from the Business Formation Statistics." *Entrepreneurship and Innovation Policy and the Economy* 1:9–42.

Hamilton, Barton H. 2000. "Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment." *Journal of Political Economy* 108, no. 3: 604–31.

Howell, Sabrina T., Josh Lerner, Ramana Nanda, and Richard R. Townsend. 2020. "How Resilient Is Venture-Backed Innovation? Evidence from Four Decades of U.S. Patenting." Working Paper 27150. Cambridge, Mass.: National Bureau of Economic Research.

Nanda, Ramana, and Matthew Rhodes-Kropf. 2013. "Investment Cycles and Startup Innovation." *Journal of Financial Economics* 110, no. 2: 403–18.

Rosenthal, Stuart S., and William C. Strange. 2012. "Female Entrepreneurship, Agglomeration, and a New Spatial Mismatch." *Review of Economics and Statistics* 94, no. 3: 764–88.

Schumpeter, Joseph A. 1943. *Capitalism, Socialism and Democracy*. Oxford: Taylor and Francis.

Sheiner, Louise, and Nasiha Salwati. 2022. "How Much Is Long COVID Reducing Labor Force Participation? Not Much (So Far)." Working Paper 80. Washington: Hutchins Center on Fiscal and Monetary Policy at Brookings.

Small, Mario L., Armin Akhavan, Mo Torres, and Qi Wang. 2021. "Banks, Alternative Institutions and the Spatial-Temporal Ecology of Racial Inequality in US Cities." *Nature Human Behaviour* 5:1622–28.

GENERAL DISCUSSION John Sabelhaus asked the authors about the role of the social safety net in a broad sense, including student loan forgiveness, for understanding business formation during COVID-19. While issues such as financing constraints tend to be front of mind when discussing start-ups, Sabelhaus noted how the risk environment for entrepreneurs changed significantly during COVID-19 and how government intervention played an important role in enabling more people to start businesses.

Related to the active role of policy during this period, Ben Harris pointed to a range of programs specifically designed to route capital to small businesses, including the $800 billion in the Paycheck Protection Plan (PPP), $10 billion passed through the American Rescue Plan for the State Small Business Credit Initiative, and $9 billion through the Emergency Capital Investment Program.[1] He asked the authors to what extent they believed the stimulus programs were part of the story.

Moritz Schularick asked the authors to speculate on the literature on aggregate demand conditions and business formation and how that relates to the fact that this period was one marked by extensive government stimulus and relief efforts.

Şebnem Kalemli-Özcan suggested linking some of the results to the broader macro picture. First, she was curious about how entrepreneurs financed their new businesses and the extent to which personal savings played a role on the backdrop of the PPP. Second, Kalemli-Özcan noted

1. US Small Business Administration, "SBA Announces Opening of Paycheck Protection Program Direct Forgiveness Portal," https://www.sba.gov/article/2021/jul/28/sba-announces-opening-paycheck-protection-program-direct-forgiveness-portal; US Department of the Treasury, "State Small Business Credit Initiative (SSBCI)," https://home.treasury.gov/policy-issues/small-business-programs/state-small-business-credit-initiative-ssbci; US Department of the Treasury, "Emergency Capital Investment Program," https://home.treasury.gov/policy-issues/coronavirus/assistance-for-small-businesses/emergency-capital-investment-program.

that, during COVID-19, labor allocation was limited and asked how this fact could be reconciled with the authors' findings.[2]

Ryan Decker agreed that trying to incorporate the effect of policy in thinking about changes in business formation is important. Related to the risk environment, Decker commented that he believed there was perhaps a greater risk appetite given the expansion of the safety net, a change in people's sense that "everything will be all right." At the same time, he was struck by how business applications in the 2021 administrative data were rising despite many of the government stimulus programs coming to an end or having already come to an end.

John Haltiwanger clarified that the PPP was not for new businesses but rather for existing businesses—and theory suggests that this may in fact stifle entry. Haltiwanger said that while there has been concern that individuals set up employer identification numbers (EINs) in order to be eligible for PPP, data from the Census Bureau matched to the Business Formation Statistics (BFS) suggest this is not the case. Thus, fraud wouldn't be able to explain the surge, either. Along the same line of argument, Haltiwanger noted that we have seen a strong labor market for two plus years now, with lots of opportunities for employment, and we still had an enormous surge in business applications.

Pinelopi Goldberg argued that as people relocate, demand for services is expected to increase in these locations, explaining some of the new business entries. Similarly, we would expect to see exit rates increase in other locations. Consequently, Goldberg was interested in exploring what the net entry rate looked like.

Ayşegül Şahin asked the authors to discuss which part of the wage distribution workers who turned self-employed came from, stating that she thought it was of importance to wage dynamics. Following up on Şahin's comment, Gerald Cohen asked if there was a way to link micro data such as the Current Population Survey (CPS) to the authors' findings, to identify educational attainment and other characteristics of the newly self-employed. That would help shed light on the extent to which the rise in new businesses would bring increases in productivity, Cohen suggested.

In terms of gaining a more detailed understanding of who these new entrepreneurs are, Haltiwanger pointed to the possibility of integrating the BFS with the Longitudinal Business Database, and with the Longitudinal

2. John Fernald and Huiyu Li, "The Impact of COVID on Productivity and Potential Output," in *Economic Policy Symposium Proceedings: Reassessing Constraints on the Economy and Policy* (Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City, 2022).

Employer-Household Dynamics data, which can provide information on who started a business and who was hired, noting that this is an important avenue for future work. He also mentioned that in joint work with Dinlersoz, Dunne, and Penciakova they found enormous spatial variation, suggesting the propensity for entrepreneurship differs by location.[3] He further praised the work of discussant Jorge Guzman focusing on racial disparities in access to finance, which Haltiwanger argued is a first-order issue.[4]

Katharine Abraham questioned the paper's implicit assumption that all employer businesses are a primary activity, arguing this need not be the case. She offered the example of a catering business, which likely would have employees but could be something a person ran on weekends. Abraham also questioned the use of CPS data for drawing conclusions about how multiple job holding has changed over time, citing known issues in those data with undercounting the number of secondary jobholders. Offering advice to the authors, Abraham suggested they could use the data employed in their study to explore what kinds of jobs new businesses have been creating. It would be interesting, for example, to know how intensive these new jobs are, something that could be proxied using payroll per added employee.

Haltiwanger agreed that the CPS data do not track self-employment well in general, as documented by Abraham and others (2018), and that it is interesting to consider both new employer and nonemployer businesses. The focus of the paper is on new employer businesses but there has also been a surge in applications for likely new nonemployers as seen in figure 1. Nonemployer businesses are important; overall there are more than 25 million nonemployer businesses, compared to a little more than 6 million employer businesses.[5] Most nonemployers are very small, but

3. Emin Dinlersoz, Timothy Dunne, John C. Haltiwanger, and Veronika Penciakova, "The Local Origins of Business Formation," working paper CES-23-34 (Washington: Center for Economic Studies, 2023).

4. Catherine E. Fazio, Jorge Guzman, Yupeng Liu, and Scott Stern, "How Is COVID Changing the Geography of Entrepreneurship? Evidence from the Startup Cartography Project," working paper 28787 (Cambridge, Mass.: National Bureau of Economic Research, 2021), https://www.nber.org/papers/w28787.

5. Katharine G. Abraham, John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer, "Measuring the Gig Economy: Current Knowledge and Open Issues," in *Measuring and Accounting for Innovation in the Twenty-First Century*, Carol Corrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel, eds. (Chicago: University of Chicago Press, 2021); US Small Business Administration, Office of Advocacy, "Frequently Asked Questions," March 2023, https://advocacy.sba.gov/wp-content/uploads/2023/03/Frequently-Asked-Questions-About-Small-Business-March-2023-508c.pdf.

nonemployers that have an EIN are larger, as discussed in the paper. Nonetheless, Haltiwanger conceded that, to Abraham's point, any new nonemployer businesses may still reflect mainly secondary activities.

In light of decreasing self-employment rates, Betsey Stevenson remarked that while we did see labor reallocation and increased entrepreneurship supported by the expansion of the social safety net during COVID-19, we ought to consider the extent to which the ability to form new businesses constitutes part of the safety net as well, helping individuals weather a storm when there are no employers around.

Robert Hall noted that a huge number of workers were placed on layoff in April 2020. Over the next few months, they were recalled to their existing jobs.[6] Hall suggested that the rapid rate of return to existing jobs is an important fact that should be kept in mind in studying business formation during this period.

To the points of Stevenson and Hall, Haltiwanger thought that there might have been a lot of brainstorming related to entrepreneurship going on, particularly in the first period of the pandemic—people wanted to do things differently, and many were not in their offices.

Martin Baily steered the discussion toward productivity and brought up the ambiguity of projected productivity at the beginning of the pandemic. While the authors suggested there was a sense of general pessimism, several sources deemed positive productivity growth likely: work by Barrero, Bloom, and Davis pointed to reallocation effects which could be positive for productivity; Goldman Sachs expected that there would be a productivity surge following some creative destruction, and McKinsey produced a study that suggested there would be increases in investment and an expansion of new technologies.[7] Baily continued, saying that in retrospect, while there were some fluctuations in productivity, the trend ultimately did not

6. Robert E. Hall and Marianna Kudlyak, "The Unemployed with Jobs and without Jobs," *Labour Economics* 79 (2022): 102244.

7. Jan Hatzius, Joseph Briggs, Devesh Kodnani, and Giovanni Pierdomenico, "The Potentially Large Effects of Artificial Intelligence on Economic Growth," Goldman Sachs, March 26, 2023, https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html; Jose Maria Barrero, Nicholas Bloom, and Steven J. Davis, "COVID-19 Is Also a Reallocation Shock," working paper 27137 (Cambridge, Mass.: National Bureau of Economic Research, 2020); Shaun Collins, Ralf Dreischmeier, Ari Libarikian, and Upasana Unni, "Why Business Building Is the New Priority for Growth," *McKinsey Quarterly*, December 10, 2020, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-business-building-is-the-new-priority-for-growth.

go anywhere, as documented by John Fernald and Huiyu Li.[8] He asked the authors whether they believed that the increase in dynamism, which they speculated about, would lead to increases in productivity.

In response to Baily's comment, Haltiwanger made the point that while new small businesses may not all turn into the next big tech firm, they do represent a form of economic mobility not just for themselves but for the workers they hire. To Şahin's point, he noted that such hires are often low-skill labor. Haltiwanger believed that ranking industries to determine where innovation will come from next provides very crude information, emphasizing that every industry has a right tail which provides important contributions to innovation, productivity, and job creation. He also pointed to recent work on the particularly high rates of entrepreneurship documented in neighborhoods with a higher proportion of Black residents as a reason for preliminary optimism and an important avenue for continued research.[9] Nonetheless, Haltiwanger highlighted that professional, scientific, and technical services have historically been particularly important for innovation—with the last productivity surge in the 1990s—but he noted that the effect from a surge in entry in high-tech sectors comes with a lag: previous research has shown that the productivity response comes six to nine years after an entry surge.[10] Consequently, we should expect that any effect on productivity this time around would also take some time to materialize. Work by Gort and Klepper, as well as by Jovanovic and MacDonald, makes a compelling argument about how entrants induce innovation.[11] But innovation also spurs new business formation—the causality goes both ways, Haltiwanger concluded.

On the topic of innovation, Michael Falkenheim wondered whether there might be lessons to be learned from the literature on war for business formation and entrepreneurship, noting that COVID-19 was a similarly destructive event, and as such may also give rise to creativity.

8. Fernald and Li, "The Impact of COVID."

9. Fazio, Guzman, Liu, and Stern, "How Is COVID Changing the Geography."

10. Lucia Foster, Cheryl Grim, John C. Haltiwanger, and Zoltan Wolf, "Innovation, Productivity Dispersion, and Productivity Growth," in *Measuring and Accounting for Innovation in the Twenty-First Century*, Carol Carrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel, eds. (Chicago: University of Chicago Press, 2021).

11. Steven Klepper, "Entry, Exit, Growth, and Innovation over the Product Life Cycle," *American Economic Review* 86, no. 3 (1996): 562–83, http://www.jstor.org/stable/2118212; Michael Gort and Steven Klepper, "Time Paths in the Diffusion of Product Innovations," *Economic Journal* 92, no. 367 (1982): 630–53; Boyan Jovanovic and Glenn M. MacDonald, "The Life Cycle of a Competitive Industry," *Journal of Political Economy* 102, no. 2 (1994): 322–47, http://www.jstor.org/stable/2138664.

Decker pondered whether the pandemic represented a persistent shock to the pace of entry; he expressed some skepticism but noted that one might need only a few cohorts of really innovative new firms in scientific and technical services in order to see an effect on productivity down the line, noting that recent entries include businesses that are helping other firms undergo technical change, such as IT consulting, engineering consulting, and data centers.

Iván Werning suggested that it may be useful to look at the outcome in other countries to perhaps gain additional insights given that we were all affected by COVID-19. Janice Eberly pointed to the United Kingdom as an example, noting that while workers were paid to stay with their employer through the Coronavirus Job Retention Scheme, many still ended up leaving.

GUIDO LORENZONI
University of Chicago

IVÁN WERNING
Massachusetts Institute of Technology

Wage-Price Spirals

ABSTRACT We interpret recent inflation experience through the lens of a New Keynesian model with price and wage rigidities and nonlabor inputs in inelastic supply. The model provides a natural interpretation of some features of the recent episode: an initial surge of noncore inflation, followed by a lagged response of core inflation and a further lagged, persistent response of wage inflation. The model also provides a natural way of discussing the role and the strength of wage-price spiral dynamics in price-setting models. The model interprets recent developments as symptoms of underlying supply constraints, which can be triggered by both demand and supply shocks. The immediate manifestation of these constraints is in the relative price of scarce, inelastic nonlabor inputs (including energy). The secondary effects arise because they produce a gap between lowered real wage aspirations of firms—that try to make up for higher nonlabor costs—and increased real wage aspirations of workers—caused by increased labor demand. The gap produces a wage-price spiral, which continues as long as the initial relative scarcity of nonlabor inputs persists, even though input prices are falling. In this view, the fact that nominal wage growth is currently exceeding price inflation can be given a benign interpretation, as a sign of real wages going back to trend and not necessarily as a concern of an ongoing spiral.

Conflict of Interest Disclosure: Guido Lorenzoni is a consultant for the Federal Reserve Bank of Chicago. The authors did not receive financial support from any firm or person for this paper or, other than the aforementioned, from any firm or person with a financial or political interest in this paper. The authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper.

Brookings Papers on Economic Activity, Fall 2023: 317–367

The recent inflation surge in the United States and in the rest of the world has reignited debates about inflation's origins and propagation mechanisms. In particular, it has brought to the forefront the separate roles and interaction of prices, wages, and profits, and indeed it has done so at two key junctures.

Early on, at the first juncture, many worried that inflation would emanate from a tight labor market, stimulated by expansionary fiscal and monetary policies, causing wage inflation that would then produce price inflation.[1] This is not how inflation played out though. Instead, price inflation and profit margins soared, while wage growth picked up later and more gradually, implying an initial fall in real wages as shown in figure 1.[2]

More recently, as price inflation started falling, wage growth rose, surpassing inflation and leading to a rise in real wages. At this second juncture, the concern is that higher wage growth would prevent inflation from going back to target, or even set off an out-of-control wage-price spiral.

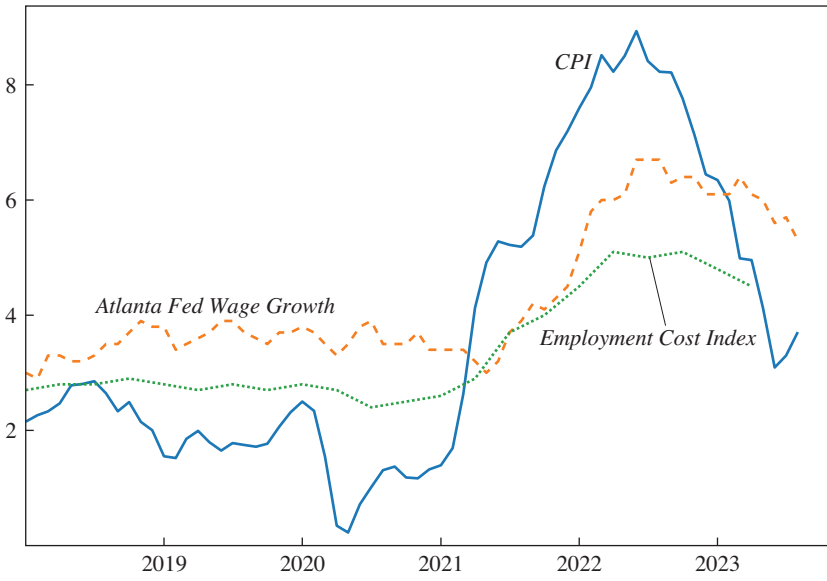
This paper aspires to simultaneously improve our understanding of these recent events while sharpening underlying economic concepts and intuitions surrounding inflation. To this end, we lay out a simple macroeconomic model. We show that this simple model is capable of capturing some key features of the recent episode. Our conceptual analysis dissects the role of prices and wages, isolating their interaction to provide a working definition of wage-price spiral and to understand the dynamics of the real wage.

Our model is relatively close to standard models, but with two essential features not always present in the most basic New Keynesian setups. One important feature of our analysis is the inclusion of a scarce nonlabor input with low substitutability in production (lower than Cobb-Douglas). We do not have in mind general forms of capital but rather inputs like energy, other primary commodities, or intermediate inputs that may be subject to shortages or in relatively fixed supply in the short run, for example, lumber or microchips. These nonlabor inputs provide both a potential supply shock or a supply constraint for demand shocks. This feature of our model is motivated by the 2020–2023 COVID-19 crises and post-COVID-19 recovery.

1. Economists who sent prescient, early warnings on inflation risk, like Blanchard (2021), focused on this transmission mechanism.

2. In the figure, along with Consumer Price Index (CPI) inflation, we show two measures of wage inflation, both of which avoid including compositional effects: the Bureau of Labor Statistics Employment Cost Index (all civilian workers, twelve-month change) and the Federal Reserve Bank of Atlanta's Wage Growth Tracker (overall, twelve-month change).

Figure 1. Post-pandemic Price and Wage Inflation in the United States



Source: Bureau of Labor Statistics and Federal Reserve Bank of Atlanta.

The other important feature of our model is that we include both nominal price and wage rigidities, as in many medium-scale models, but unlike the simplest New Keynesian models with only one form of nominal rigidity.

In a model with these features, supply constraints play a crucial role in inflation dynamics, and when these supply constraints are active, both demand and supply disturbances can set in motion price and wage dynamics that resemble the ones observed.

Namely, the model can produce a three-phase pattern of adjustment in nominal prices. First, there is a bout of very high inflation in the price of the inelastic nonlabor inputs, followed by a prolonged gradual fall in the price of these inputs. Second, there is a more persistent period of high general good price inflation. Third, there is a smaller but even more persistent increase in wage inflation.

The pattern described follows from our assumptions on the role of the inelastic input, which more directly affects price-setting firms, and on the relative degree of price stickiness with the input price being perfectly flexible and goods prices being more flexible than wages. This pattern implies that, at some point, wage inflation crosses price inflation, so a period in which real wages fall is followed by a period in which they recover.

Data are always interpreted with a theoretical lens. At one end of the spectrum, commentators and Federal Reserve governors' speeches often employ standard macroeconomic concepts, such as a Phillips curve, in their simplest incarnations, to fix ideas or make back-of-the-envelope calculations. On the other end of the spectrum, several papers have contributed by calibrating sophisticated multi-sectoral models. Our paper lives in the gap between these extremes—our model is simpler than medium scale calibrated models, allowing us to develop several important concepts, yet it goes beyond textbook tools used in day-to-day policy debates.

Turning to the more conceptual points of our paper, one may ask, what do we mean by a wage-price spiral? While there may not be universal agreement, in this paper we use the expression to describe a feedback mechanism where wages and prices compete adjusting upward: wage earners try to keep up with rising prices; price setters try to keep up with rising wages. This mechanism amplifies and perpetuates the effects of certain inflationary shocks.

Our perspective is that this feedback mechanism is present in virtually all models—including standard New Keynesian varieties. The purpose of this paper is to elucidate and explore this mechanism in detail and focus on the shape of price and wage responses to both supply and demand shocks.

At heart, the economic logic of the wage-price spiral mechanism is that workers and firms disagree on the relative price of goods and labor, that is, on the real wage W/P. When firms adjust nominal prices, they do so with some goal for W/P. But workers may have a different, higher goal for W/P and set nominal wages to reach that goal. If they do, the outcome of this disagreement is nominal escalation, with inflation in both prices and wages.

Our interpretation of the concept of a wage-price spiral, highlighting disagreement or conflict as a proximate cause of inflation, is an idea that we explore more generally in Lorenzoni and Werning (2022). The present paper studies how this conflict plays out in particular variants of the New Keynesian model and places attention on the path of real wages in response to demand and supply shocks.

Beyond providing an interpretation of recent inflation dynamics, we also use our model to derive a number of general positive and normative results.

First, we derive a general condition for the direction of adjustment of the real wage in response to demand shocks. We show that whether the real wage increases or decreases following a demand shock depends on how strong the forces set in motion on the price-setting side of the model and on the wage-setting side are.

A demand shock acts on the price side by producing an endogenous increase in the price of nonlabor inputs. If there is low degree of substitutability between labor and nonlabor inputs, we get both a large price response of nonlabor inputs and a large reduction in the marginal product of labor when nonlabor inputs are relatively scarce. The first force will show up in noncore inflation measures. The second will contribute to a distributional tension between workers and firms that materializes in a wage-price spiral.

A demand shock also acts on the wage side directly. Our model does not feature unemployment and search directly, but the labor supply side of our model captures the basic idea that an overheated labor market will directly affect nominal wage demands by increasing the rate at which workers are willing to exchange labor for consumption goods. Therefore, this piece of the model captures the basic logic of a wage Phillips curve. Through this channel, excess demand will also produce higher real-wage aspirations for workers and contribute to the wage-price spiral.

Excess demand operates and contributes to a wage-price spiral on *both* sides. However, for the movement in the real wage, what matters is the relative strength on the two sides. In our low-elasticity-of-substitution calibration, the effect is stronger on the price side and thus produces overall lower real wages.[3]

An additional observation that comes from our analysis is that both demand and supply shocks create a situation of excess demand. In the demand shock case, natural output is unchanged, but the demand temporally expands. In the supply shock case, the "natural" level of output is lower, but the demand is unchanged. This excess demand leads to a tension between the level of the real wage that firms and workers aspire to, resulting in a wage-price spiral that produces inflation in both wages and prices. However, excess demand is not a sufficient statistic. In the supply shock case, real wages always fall; whereas in the demand shock case, the real wage may fall depending on parameters. Only under some conditions are the effects on wages and prices similar for both shocks.

Excess demand is zero when there is a zero output gap. A result that applies in our model is that, with a zero output gap, there can never be both

3. Incidentally, our analytical result can be taken as a contribution to the classic debate on the cyclicality of the real wage that has spurred a large body of literature, including Christiano and Eichenbaum (1992) and Rotemberg and Woodford (1992). However, our aim here is not to discuss the general cyclical property of real wages but rather to discuss how potentially sizable real wage movements can be set in motion in special circumstances, like the recent post-pandemic recovery.

price and wage inflation, that is, price and wage inflation always have the opposite sign. Furthermore, our definition of conflict inflation (Lorenzoni and Werning 2022), which we use to capture the wage-price spiral force, is closely related to the size of the output gap in the New Keynesian model here. This connects us immediately to the notion of "divine-coincidence inflation" introduced by Rubbo (2020), which in the model here coincides with conflict inflation.[4]

The result just stated can be rephrased as that if the central bank successfully pursues a zero output gap, the central bank can always prevent a wage-price spiral (i.e., achieve zero conflict inflation). But it does not imply that a zero output gap policy is the optimal policy. In section IV, we study optimal policy and ask two questions. First, could it be part of optimal policy to "run the economy hot," that is, to allow for a positive output gap despite high inflation? Second, could it be part of optimal policy to go further and allow for inflation in both prices and wages?

Our answer to the first question is affirmative: if the economy needs a lower real wage, it may be more efficient to reach the adjustment with the help of higher price inflation and moderate wage deflation, rather than through lower price inflation and deeper wage deflation. A positive output gap helps shift the adjustment in the direction of price inflation, so it is socially beneficial in this manner.

The answer to the second question is also affirmative. We construct examples in which, at some point along the adjustment path, the output gap is positive, and price and wage inflation are both positive. The economic intuition is that this aspect of policy is a form of "forward guidance": by promising to heat up the economy in the future, we speed up the adjustment of the real wage today. Underlying this result is the assumption of forward-looking price- and wage-setting behavior and the commitment of policy. In contrast, when policy has full discretion, the equilibrium outcome never features both price and wage inflation.

There is a large and growing body of literature analyzing the post-pandemic surge in inflation in the United States and globally. Our paper is part of a group of papers that emphasizes the crucial role of supply disruptions and supply constraints in the recent inflation surge, a group that includes Ball, Leigh, and Mishra (2022), Amiti and others (2023), Bernanke and Blanchard (2023), Comin, Johnson, and Jones (2023), Gagliardone and Gertler (2023), and Kabaca and Tuzcuoglu (2023). We do it here by pointing

4. This is connected to the "divine-coincidence" inflation index of Rubbo (2020), which also only depends on the output gap.

out the explanatory power of this interpretation for the joint dynamics of prices and wages.

The way in which supply constraints play out here is closely related to the approach in Comin, Johnson, and Jones (2023), who develop a quantitative model with an explicit treatment on nonlinearities in the supply of nonlabor inputs and take an explicit open economy approach. We believe the virtue of this way of interpreting the facts is that it shows a state of global excess demand can cause endogenously sharp input price adjustments, which cannot be taken merely as exogenous price shocks.

Our model emphasizes the role of the real wage as a state variable. This plays an important role in our interpretation of recent events. In particular, we see the recent increase in the real wage as fundamentally driven by a desire of wage setters to make up for the accumulated losses in purchasing power during the early stage of the episode. In other words, we interpret the recent high wage inflation as driven by some form of catch-up. The empirical analysis by Bernanke and Blanchard (2023) provides an empirical challenge to this view, as they attempt to measure this catch-up mechanism in the data and fail to find it significant. However, it is not easy to identify structurally this channel of catch-up, and in general, findings of wage inflation responding to past price inflation can be taken as supportive of a lag effect, leading to a lag recovery of real wages.[5]

In terms of the broader idea of wage-price spiral, our paper is connected to a vast amount of literature, and we will make only a few close references here. Blanchard (1986) wrote the seminal paper connecting that idea to New Keynesian models of staggered price setting. The model has nominal prices and wages that are fixed for two periods, with prices reset in even periods and wages in odd periods. The main result in the paper is that the alternating wage and price setting leads to a slow adjustment of the price level in response to a permanent money supply shock and the adjustment features dampening oscillations in the real wage. Our paper instead builds on the canonical New Keynesian setting with sticky price and sticky wages of the Calvo variety as developed by Erceg, Henderson, and Levin (2000). Relative to Blanchard (1986), price and wage setting occur in a staggered fashion without the predictable alternation between wages and prices, so our model is not prone to the same type of oscillations. We also do not focus on a permanent money shock or study monetary policy in terms of money supply. Instead, we focus on supply and demand shocks under different policy responses. Finally, we investigate optimal monetary policy.

5. See, for example, the regressions in Barlevy and Hu (2023) and the literature cited there.

Our analysis of wage-price spirals in section II builds on the idea of inflation as the result of distributional conflict, something we explore in more detail in Lorenzoni and Werning (2022). A seminal contribution on this conflict perspective of inflation is Rowthorn (1977). That paper provides a model where, in each period, wages are first set by workers and then prices are set by firms. Inflation is shown to be increasing in the conflict or "aspirational gap." Because of the assumed sequential timing of price and wage setting, conflict and inflation must not be fully anticipated by workers. Indeed, no rational expectations equilibrium exists with conflict. In contrast, our model features staggered wages and prices, which ensure that there is an equilibrium with finite conflict and inflation, even under rational expectations.

Our modeling of nonlabor inputs and their connection to price and wage determination connects our analysis to extensive literature on models of energy shocks.[6] An important modeling difference is that we focus on nominal wage rigidities, while they study a form of real-wage rigidity.

On the normative side, our paper is connected to the welfare analysis of alternative policy rules in models where both prices and wages are rigid, going back to the original paper by Erceg, Henderson, and Levin (2000) and to the real rigidity model by Blanchard and Galí (2007b). The starting observation in the literature is that the presence of both price and wage rigidities breaks divine coincidence and introduces potentially interesting trade-offs in the response of monetary policy to supply shocks. We offer a complete characterization of optimal policy and explore conditions for the optimum to have a positive output gap in combination with high inflation, as well as cases where it is optimal to have both wage and price inflation.

I. Model

We build our arguments in a standard New Keynesian model with nominal price and wage rigidities. To capture supply shocks, an important ingredient we include is a scarce nonlabor input X, which is used alongside labor for production. We assume this input has a flexible price, and we allow the production function to have elasticity of substitution different from one.[7] An important example is energy inputs, but we interpret X more broadly to

6. For example, Blanchard and Galí (2007a); in turn, this connects us to the enormous body of literature on the effects of oil shocks, going back to Bruno and Sachs (1985).

7. This is formally equivalent to having labor and capital, with capital rented at a flexible price, although the interpretation is different. Erceg, Henderson, and Levin (2000) have labor and capital. Closer to the interpretation here, Blanchard and Galí (2007a) have an energy input.

also capture shortages, bottlenecks, and capacity constraints in the supply of intermediates like microchips or lumber, which have been in the spotlight during the post-pandemic recovery.

We focus on a closed economy in which the supply of X is given while the price of X adjusts endogenously in equilibrium. The analysis can be easily expanded to the case of an open economy in which the good X is imported, and, in particular, to the limited case of a small open economy that takes the world price of X as given. In that case, a supply shock would take the form of a shock to the world price instead of a shock to the endowment.

I.A. Setup

Time is continuous and infinite. The representative household has preferences

$$\int_0^\infty e^{-\rho t}\left(\frac{1}{1-\sigma}C_t^{1-\sigma}-\frac{\Phi_t}{1+\eta}N_t^{1+\eta}\right)dt,$$

where C_t is an aggregate of a continuum of varieties of goods

$$C_t=\left(\int_0^1 C_{jt}^{1-\frac{1}{\varepsilon_C}}dj\right)^{\frac{1}{1-\frac{1}{\varepsilon_C}}},$$

N_t is labor supply, and Φ_t is a labor supply shock. Each goods variety j is supplied by a monopolistic firm with production function

$$Y_{jt}=F\left(L_{jt},X_{jt}\right)\equiv\left(a_L L_{jt}^{\frac{\epsilon-1}{\epsilon}}+a_X X_{jt}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}},$$

where L_{jt} is the labor input and X_{jt} is the nonlabor input. The labor input L_{jt} of each firm j is an aggregate of a continuum of labor varieties

$$L_{jt}=\left(\int_0^1 L_{jkt}^{1-\frac{1}{\varepsilon_L}}dk\right)^{\frac{1}{1-\frac{1}{\varepsilon_L}}}.$$

Each labor variety k is supplied by a monopolistic union that employs labor from households and turns it, one for one, into specialized labor services of type k. Integrating over firms, total employment of labor variety k is

$N_{kt} = \int_0^1 L_{jkt} dj$. Integrating over unions, total labor supply is $N_t = \int_0^1 N_{kt} dk$. The representative household owns an exogenous endowment X_t of the nonlabor input X and sells it to the monopolistic goods producers on a competitive market, at the price P_{Xt}.

Monopolistic firms set the nominal price at which they are willing to sell their variety and then supply the amount chosen by consumers. Similarly, monopolistic unions set the nominal wage and supply the amount chosen by firms. Firms and unions are only allowed to reset their price and their wage rate occasionally. Namely, at each point in time, firms are selected randomly to reset their price with Poisson arrival λ_p, and unions are selected with arrival λ_w.

When the exogenous variables X_t and Φ_t are constant, the model has a steady state in which quantities are constant, nominal prices are constant (zero inflation), all goods varieties have the same price, and all labor varieties have the same wage. We will consider an economy in steady state and analyze its response to one-time, unexpected shocks, either due to changes (transitory or permanent) to X_t or Φ_t, or to changes in monetary policy leading to transitory deviations of C_t and N_t from the path consistent with zero inflation.

I.B. Price and Wage Setting

Let P_t^* and W_t^* denote the price and wage set by the firms and unions that can reset at time t, while P_t and W_t denote the price indexes for the goods and labor aggregates.

The nominal marginal cost of producing good j is

$$\frac{W_t}{F_L\left(L_{jt}, X_{jt}\right)} = \frac{W_t}{a_L Y_{jt}^{\frac{1}{\epsilon}} L_{jt}^{-\frac{1}{\epsilon}}}.$$

Using lowercase variables to denote log-linear deviations from steady state and taking a first-order approximation, nominal marginal costs can then be expressed as

$$w_t - mpl_{jt}, \tag{1}$$

where

$$mpl_{jt} = \frac{1}{\epsilon}\left(y_{jt} - l_{jt}\right)$$

is the marginal product of labor. The production function of firm j in log-linear approximation is

$$y_{jt} = s_L l_{jt} + s_X x_{jt}, \tag{2}$$

where s_L and s_X are the steady-state shares of the labor and nonlabor inputs, with $s_L + s_X = 1$. All firms being price takers in the input market, they all employ inputs in the same ratio L_{jt}/X_{jt}, so in log-linear approximation

$$l_{jt} - x_{jt} = n_t - x_t,$$

where n_t and x_t are the aggregate supplies of the two inputs. Combining these results, the marginal product of labor is

$$mpl_t = \frac{s_X}{\epsilon}\left(x_t - n_t\right). \tag{3}$$

Following standard steps, optimal price setting requires that firms set their price at time t equal to an average of future nominal marginal costs, conditional on not resetting. This gives the following optimality condition for P_t^* in log-linear approximation:

$$p_t^* = \left(\rho + \lambda_p\right)\int_t^\infty e^{-(\rho+\lambda_p)(\tau-t)}\left(w_\tau - mpl_\tau\right)d\tau. \tag{4}$$

Following similar steps, we can derive the wage-setting equation

$$w_t^* = \left(\rho + \lambda_w\right)\int_t^\infty e^{-(\rho+\lambda_w)(\tau-t)}\left(p_\tau + mrs_\tau\right)d\tau, \tag{5}$$

where

$$mrs_t = \phi_t + \sigma y_t + \eta n_t \tag{6}$$

is the marginal rate of substitution between consumption and leisure of the representative consumer.

The presence of w_τ on the right-hand side of equation (4) and p_τ on the right-hand side of equation (5) captures the logic of a wage-price spiral in our model. Firms aim to get prices to be a constant markup over nominal marginal costs, and since marginal costs depend on nominal wages, they set nominal prices to catch up with current and anticipated future nominal

wages. Symmetrically, wage setters aim to achieve a real wage that reflects their willingness to substitute leisure with consumption goods, so they set nominal wages to catch up with current and anticipated future nominal goods prices.

The optimality condition for the input ratio of firms can be written as follows:

$$p_{Xt} = w_t - \frac{1}{\epsilon}\left(x_t - n_t\right). \tag{7}$$

This condition will be used to derive the equilibrium input price p_{Xt}.

I.C. Inflation Equations

To go from equations (4) and (5) to wage and price inflation, combine them with the differential equations for p_t and w_t:

$$\dot{p}_t = \lambda_p\left(p_t^* - p_t\right) \text{ and} \tag{8}$$

$$\dot{w}_t = \lambda_w\left(w_t^* - w_t\right). \tag{9}$$

As shown in the online appendix, we then obtain the following expressions:

$$\rho\pi_t = \Lambda_p\left(\omega_t - mpl_t\right) + \dot{\pi}_t \text{ and} \tag{10}$$

$$\rho\pi_t^w = \Lambda_w\left(mrs_t - \omega_t\right) + \dot{\pi}_t^w, \tag{11}$$

where we use the notation $\pi_t \equiv \dot{p}_t$ and $\pi_t^w \equiv \dot{w}_t$ for price and wage inflation and $\omega_t \equiv w_t - p_t$ for the real wage, and the coefficients Λ_p and Λ_w are

$$\Lambda_p = \lambda_p\left(\rho + \lambda_p\right) \text{ and } \Lambda_w = \lambda_w\left(\rho + \lambda_w\right).$$

Real wage dynamics are given by

$$\dot{\omega}_t = \pi_t^w - \pi_t. \tag{12}$$

Equations (10) and (11) can be interpreted in terms of a conflict between the real wage aspirations of workers and firms, an interpretation we develop

in Lorenzoni and Werning (2022). In the context of the New Keynesian model, the workers' aspiration is given by the marginal rate of substitution mrs_t at which the representative worker is willing to exchange labor for goods, and the firms' aspiration is the marginal product of labor mpl_t.[8] As in Lorenzoni and Werning (2022), a discrepancy between the aspirations mpl_t and mrs_t is the proximate cause of inflation.

Equations (10) and (11) can also be expressed as traditional Phillips curves because the expressions $\omega_t - mpl_t$ and $mrs_t - \omega_t$ can be written in terms of gaps between equilibrium objects and their "natural" level. Focusing on the wage equation, we can write

$$\begin{aligned} mrs_t - \omega_t &= mrs_t - mrs_t^* - \left(\omega_t - \omega_t^*\right) \\ &= \left(\sigma s_L + \eta\right)\left(n_t - n_t^*\right) - \left(\omega_t - \omega_t^*\right), \end{aligned}$$

where ω_t^* is the flexible-price wage rate and n_t^* is the natural level of employment.[9] Substituting this expression in equation (11) we obtain a wage Phillips curve that connects wage inflation to the employment gap $n_t - n_t^*$. An analogous derivation can be done for the price equation. The crucial observation here is that in both Phillips curves there is an additional term, given by the deviation between the real wage and its flexible-price level ω_t^*. Notice that ω_t is a state variable of our system because both w_t and p_t move only gradually due to stickiness—at a given moment in time ω_t given by the history of past shocks.

Given an initial condition ω_0 and given paths for mpl_t and mrs_t for $t \geq 0$, the three equations (10)–(12) give unique paths for price and wage inflation.

Our approach in the rest of the paper is to split the analysis into two steps: (1) from the paths for fundamental shocks and aggregate real activity derive the paths of mpl_t and mrs_t; and (2) from the paths of mpl_t and mrs_t derive inflation. In general, in a full-blown general equilibrium model, the paths of mpl_t and mrs_t are endogenous and this way of splitting the analysis is somewhat artificial. However, a central point of this paper is to show that this decomposition helps understand the mechanisms underlying inflation in equilibrium.

8. The variable ϕ_t in the notation of Lorenzoni and Werning (2022) corresponds to mpl_t here and the variable γ_t corresponds to mrs_t.

9. This derivation applies because at the natural allocation the real wage is equalized to the workers' *mrs*. The detailed derivations are in the online appendix.

The next section focuses on step 2. We then go back to step 1 in the following section.

II. From Aspirations to Inflation, with and without a Spiral

In general, shocks to the economy translate into endogenous changes in the variables *mpl* and *mrs*, which, as argued above, reflect the real wage aspirations of firms and workers. In this section, we take the paths of *mpl* and *mrs* as given and focus on deriving inflation as a function of them. This part of the analysis isolates how staggered price setting produces inflation for given aspirations and allows us to identify the wage-price spiral mechanism. The next section shows how shocks and policies determine *mpl* and *mrs* and thus completes the analysis. A reader mostly interested in our interpretation of the post-pandemic high inflation episode can skip this section without loss.

Throughout the paper, we mostly focus on exponentially decaying paths of *mpl* and *mrs* that take the following form.[10] Before $t = 0$, the economy is in steady state: all variables expressed in log deviations from the steady state are equal to zero. At $t = 0$, there is an unexpected shock and mpl_0 and mrs_0 jump discretely to values different from zero (at least for one of them). From then on, they both converge back to the original steady state at constant speed δ, so $mpl_t = mpl_0 e^{-\delta t}$ and $mrs_t = mrs_0 e^{-\delta t}$. The demand and supply shocks analyzed in the next section produce paths with this shape, so the analysis here will immediately apply.

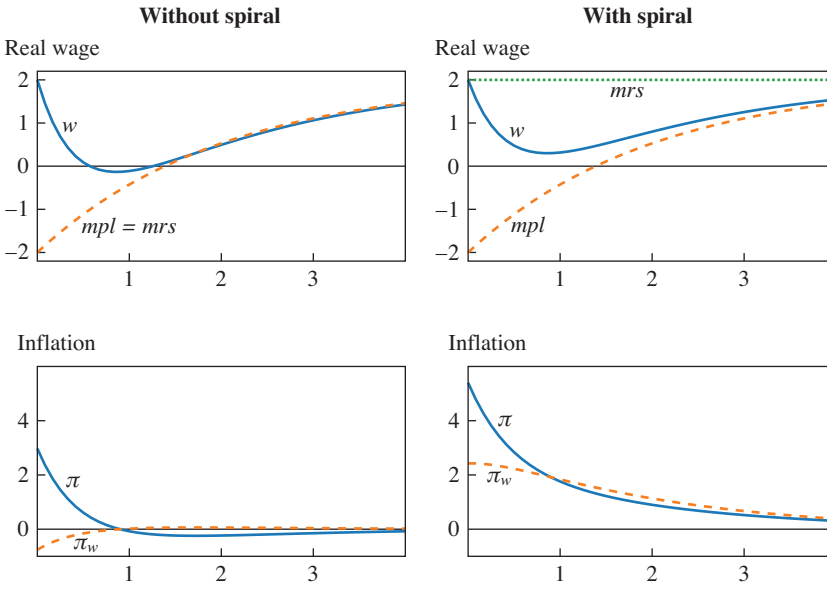
Deriving price and wage inflation from equations (10) and (11) requires solving first the endogenous path of the real wage ω_t. In other words, as mentioned earlier, the real wage is a necessary state variable in our inflation equations. The solution for the real wage in terms of *mpl* and *mrs* comes from solving a second-order ordinary differential equation (ODE); the details are provided in the online appendix. Once we have ω_t, equations (10) and (11) can be solved forward to get

$$\pi_t = \Lambda_p \int_0^\infty e^{-\rho s}\left(\omega_s - mpl_s\right)ds \text{ and} \tag{13}$$

$$\pi_t^w = \Lambda_w \int_0^\infty e^{-\rho s}\left(mrs_s - \omega_s\right)ds. \tag{14}$$

10. In the online appendix, we provide a general analytical characterization of the relation between the paths $\{mpl_t, mrs_t\}$ and price and wage inflation.

Figure 2. Aspirations and Inflation, with and without a Spiral



Source: Authors' calculations. The parameters for the examples are $\lambda_p = 2$, $\lambda_w = 1$, $\rho = 0.04$, $\delta = 0.5$.

Price and wage inflation are driven by current and anticipated gaps between the real wage and firms' and workers' aspirations. These two equations are used to provide intuition in this section.

II.A. Two Examples

Consider the two numerical examples plotted in figure 2.

In the first, *mpl* and *mrs* fall by the same amount at date 0, that is, $mrs_0 = mpl_0 < 0$. On impact, the reduction in *mpl* increases firms' marginal costs, leading firms to increase nominal prices, while the reduction in *mrs* lowers workers' aspirations and workers reduce nominal wages. In the top left panel of figure 2, we see that this leads to $\pi_0 > 0 > \pi_0^w$. The real wage starts falling, as shown in the lower left panel. As time goes by, the force of the initial shock goes away while, at the same time, the real wage is lower. Both forces reduce $\omega - mpl$ in the price inflation equation and increase $mrs - \omega$ in the wage inflation equation: the gap between aspirations and the real wage fall for both. After some date, when *mpl* and *mrs* are small enough and the real wage has fallen enough, both inflation rates π_t and π_t^w flip sign and we have $\pi_t < 0 < \pi_t^w$. From then on, the real wage starts growing and converges back to its initial level.

In this example, even though wage setters and price setters respond to each other's prices (current and anticipated), this does not produce generalized inflation or deflation, because the two parties are aiming to achieve the same relative price adjustment, so their actions tend to dampen each other. The fact that firms increase prices tends to remove the deflationary impulse on the workers' side. The fact that workers lower their wages tends to remove the inflationary impulse on the firms' side. In this case a wage-price spiral is not present.

In the second example, only the aspirations of firms change, with $mpl_0 < 0$, but mrs_0 is unchanged at zero. In this case there is a positive gap $mrs_0 - mpl_0$. This case is illustrated in the two panels on the right in figure 2.

On impact, the reduction in *mpl* increases firms' marginal costs as in the first example. Now there is no direct effect of *mrs* on the workers' side; workers anticipate a future reduction in real wages and react at date zero by raising their nominal wage demand.[11] Therefore, we get both wage and price inflation, $\pi_0 > \pi_0^w > 0$. In general, in every case where there is a unilateral change in *mpl*, with no change in *mrs*, it is possible to show that price inflation is larger than wage inflation at $t = 0$, given that the price equation is affected directly by the change in *mpl*, while the wage equation is only affected indirectly through the future equilibrium adjustment in ω.[12]

Notice the back and forth between price and wage inflation that amplifies the initial shock. The shock originates in the inflation equation but produces an undesirable relative price adjustment for workers, creating a positive gap between workers' aspirations and the real wage path, inducing wage setters to respond. This causes price inflation to spill over into wage inflation. The wage setters' response in turn dampens the adjustment in the real wage, relative to what happens in our first example: comparing the two lower panels in figure 2, the real wage ω_t falls less in the panel on the right. Therefore, the presence of wage inflation, slowing the fall in real wages, reinforces the price inflation response as firms, anticipating a weaker reduction in real wages, keep price inflation higher.[13]

11. In equation (14), $mrs_s = 0$ and $\omega_s < 0$ for all s. Why the real wage falls in this example is explained below.

12. See proposition 5 in the online appendix.

13. If nominal wages were perfectly sticky, this amplification would not be present and price inflation would be lower throughout. We go back to the relation between stickiness and amplification at the end of this section.

The expression "wage-price spiral" is used to describe these mutually reinforcing dynamics between price and wage inflation. In the first example there is no wage-price spiral, in the second there is.

II.B. Spiral Dynamics and Conflict Inflation

In the two examples above, we just argued that the first example shows no spiral while the second does. But how can we distinguish more formally the spiral force in the second from the relative price adjustment mechanism that drives nominal prices and wages in the first?

The crucial difference is that in the second example, the attempt of each side to move the relative price in its preferred direction leads to a protracted period of high inflation in both prices and wages. Let us measure the spiral effect in terms of the cumulated effect on price and wage inflation over the entire episode. Since the real wage always mean reverts to zero and cumulated price and wage inflation are the same, we can define

$$\Pi^{Spiral} \equiv \int_0^\infty \pi_t dt = \int_0^\infty \pi_t^w dt.$$

In the online appendix, we prove that

$$\Pi^{Spiral} = \frac{\Lambda_p \Lambda_w}{\Lambda_p + \Lambda_w} \frac{1}{\delta(\rho + \delta)} \left(mrs_0 - mpl_0 \right).$$

Notice the symmetric role of Λ_p and Λ_w in this expression: For the spiral effect to be present, we need prices and wages to respond to each other. If one side has fixed nominal prices, for example $\lambda_w = 0$, then the spiral is completely absent. On the other hand, if we vary λ_p and λ_w and hold fixed the total degree of nominal rigidity in the economy $\lambda_w + \lambda_p$, then the maximum power of the spiral arises when $\lambda_p = \lambda_w$, that is, when each side responds to the other with equal speed.

The spiral measure just introduced, immediately connects spiral dynamics to the notion of conflict inflation proposed in Lorenzoni and Werning (2022), which is defined as follows:

$$\Pi_t^{Conflict} \equiv \frac{\Lambda_p \Lambda_w}{\Lambda_p + \Lambda_w} \int_0^\infty e^{-\rho s} \left(mrs_{t+s} - mpl_{t+s} \right) ds,$$

and with exponentially decaying shocks, yields

$$\Pi_t^{Conflict} = \frac{\Lambda_p \Lambda_w}{\Lambda_p + \Lambda_w} \frac{1}{\rho + \delta}\Big(mrs_0 - mpl_0\Big).$$

We then conclude that

$$\Pi^{Spiral} = \frac{1}{\delta}\Pi_0^{Conflict},$$

which means that conflict inflation at date zero fully captures the underlying forces that lead to a protracted period of joint price and wage inflation.

Notice that from equations (13) and (14), we get

$$\Pi_0^{Conflict} = \alpha\pi_0 + \Big(1 - \alpha\Big)\pi_0^w, \tag{15}$$

where α is a coefficient of relative stickiness, defined as

$$\alpha \equiv \frac{\dfrac{1}{\Lambda_p}}{\dfrac{1}{\Lambda_p} + \dfrac{1}{\Lambda_w}}.$$

We then have a "forecasting" interpretation of the result above. Consider an econometrician who does not observe the underlying shocks mrs_0 and mpl_0 at $t = 0$ but only the current inflation rates π_0 and π_0^w. Conflict inflation is the linear combination of π_0 and π_0^w that provides the best estimate of the cumulated future effect of the underlying shocks on inflation.[14]

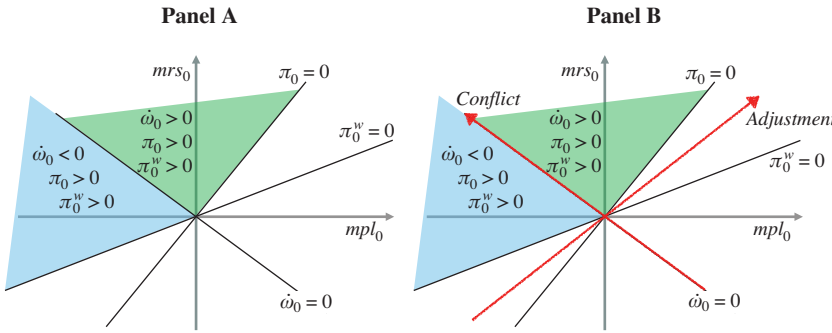
From equation (15) and $\dot{\omega}_0 = \pi_0^w - \pi_0$, we get the decomposition

$$\pi_0 = \Pi_0^{Conflict} - \Big(1 - \alpha\Big)\dot{\omega}_0 \text{ and}$$

$$\pi_0^w = \Pi_0^{Conflict} + \alpha\dot{\omega}_0.$$

14. This result relies on the simple joint AR(1) structure of the shocks to mrs_0 and mpl_0. It is an open important question how to extend the connection between conflict inflation and inflation forecasting to richer structures.

Figure 3. Regions for mpl_0 and mrs_0



Source: Authors' calculations.

Conflict inflation captures the underlying common component of price and wage inflation due to the gap between the aspirations on the two sides of the market (mpl_0 and mrs_0). The presence of the gap is crucial to set in motion mutually reinforcing responses on the two sides. When there is no gap, there can be no generalized inflation, π_t and π_t^w have opposite sign, and the mutual responses tend to dampen the initial shock, consistent with our first example.

Notice that in the New Keynesian model considered here, conflict inflation $\Pi_t^{conflict}$ is proportional to the output gap as we shall see in the next section. This implies that conflict inflation coincides with the notion of divine-coincidence inflation in Rubbo (2020) and with the composite inflation index in the optimal inflation-targeting rule of Giannoni and Woodford (2005).[15]

A GRAPHICAL REPRESENTATION A graphical representation can help interpret the decomposition above.

In panel A of figure 3 we divide the space (mpl_0, mrs_0) into six regions, depending on the sign of the three variables π_0, π_0^w, and $\dot{\omega}_t$.

Proposition 1 shows that the configuration in figure 3 is general and independent of parameters, given exponentially decaying shocks. The proposition gives conditions in terms of the coefficient ψ, which is a function of the parameters Λ_p, Λ_w, ρ, and δ, and defined in the online appendix.

15. See chapter 6, section 4 of Galí (2015) for a textbook discussion.

PROPOSITION 1. Given exponentially decaying paths for *mpl* and *mrs*, at date $t = 0$, price and wage inflation satisfy

$$\pi_0 > 0 \text{ iff } \left(1-\alpha\right)\psi \cdot mrs_0 > \left(1-\alpha\psi\right) \cdot mpl_0,$$

$$\pi_0^w > 0 \text{ iff } \left(1-\left(1-\alpha\right)\psi\right) \cdot mrs_0 > \alpha\psi \cdot mpl_0,$$

and

$$\pi_0^w - \pi_0 = \dot{\omega}_0 > 0 \text{ iff } \alpha mpl_0 + \left(1-\alpha\right)mrs_0 > 0.$$

The slope of the boundary of the $\pi_0 > 0$ region is always steeper than that of the $\pi_0^w > 0$ region.

The shaded regions in figure 3 are those in which the economy features positive price and wage inflation. Both $mrs_0 > 0$ and $mpl_0 < 0$ are inflationary forces and produce inflation as long as one of them is present and strong enough.

A positive value for mrs_0 acts directly on wage inflation, a negative mpl_0 acts directly on price inflation. Both also act indirectly through their effects on ω_t. A high mrs_0, by pushing future real wages up, tends to increase expected marginal costs and price inflation at $t = 0$. A low mpl_0, by pushing future real wages down, tends to increase wage demands and wage inflation at $t = 0$. The fact that *mrs* acts directly on wages, while *mpl* acts directly on prices gives some intuition for why the slope of the $\pi_0 = 0$ line is steeper than that of the $\pi_0^w = 0$ line.

The difference between the two shaded regions is that in the region to the left, the real wage falls at $t = 0$ while it increases in the region to the right. The reason for the difference is the relative strength of the pressure on price setters and wage setters.

Panel B of figure 3 is identical to panel A but adds two axes that represent the conflict and adjustment components of inflation.

The adjustment axis is simply given by the 45 degree line, $mrs_0 = mpl_0$, given that along that line conflict inflation is zero.

The conflict axis is the boundary between the shaded regions: it is the locus where the power of a wage-price spiral is stronger because the aspirations of workers and firms are opposite and of equal force once we adjust for the frequency of price adjustment, that is, where

$$\left(1-\alpha\right)mrs_0 = -\alpha mpl_0.$$

Along that locus there is zero adjustment inflation: the opposite efforts of workers and firms produce no movement in the real wage and only socially wasteful price dispersion.[16]

To clarify the connection between the figure and the analysis above, it is useful to remember that the figure only shows the impact effect on π_0 and π_0^w. As time goes by and ω_t changes, the same figure applies but with the origin of the conflict and adjustment axes (and of the $\pi_t = 0$ and $\pi_t^w = 0$ loci) shifting along the 45 degree line. So, for example, we can have a shock in the upper-right quadrant that initially produces $\pi_0 < 0$ and $\pi_0^w > 0$, but also gives positive conflict inflation $\Pi_0^{Conflict} > 0$. As time goes by, we will have $\omega_t > 0$ and the origin will shift to the right along the 45 degree line while, at the same time, mrs_t and mpl_t move linearly toward the (0, 0) origin. This will at some point produce a combination $\pi_t > 0$ and $\pi_t^w > 0$, consistent with the fact that the shock will eventually produce positive cumulated inflation in both prices and wages.[17]

II.C. Stickiness and Amplification

Consider now a different exercise: fix the size of two initial shocks $mrs_0 > 0$ and $mpl_0 < 0$ and change the economy's parameters λ_w and λ_p to vary the degree by which the shocks get amplified through the wage-price responses.

As we increase the speed at which either prices or wages are reset, the wage-price spiral mechanism gets stronger. This is shown in figure 4, where we plot level curves for π and π_w. The relatively steeper curves (in absolute value) correspond to π, the flatter ones to π_w. A higher frequency

16. Projecting any point (mpl_0, mrs_0) on the two axes, the conflict coordinate gives conflict inflation Π_0, while the adjustment coordinate gives $\dot{\omega}_0$. The two coordinates measure adjustment and conflict inflation if we scale the axes as follows: on the adjustment axis, the unit vector is

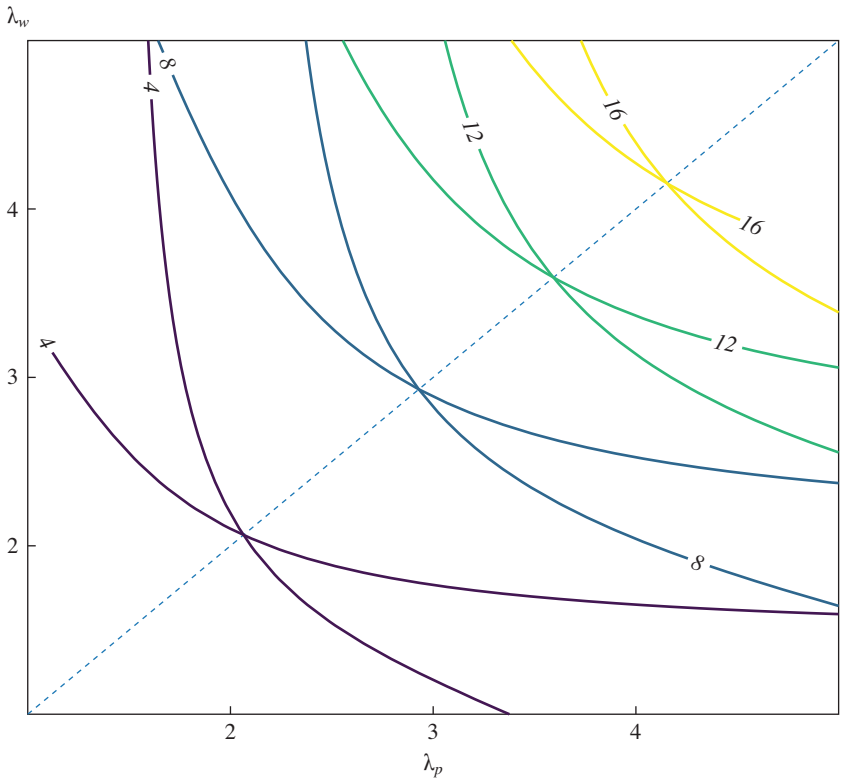
$$\begin{pmatrix} mpl_0 \\ mrs_0 \end{pmatrix} = \frac{r_2 + \delta}{\Lambda_p + \Lambda_w} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where r_2 is the positive eigenvalue of the real wage ODE, as defined in the online appendix; and on the conflict axis, the unit vector is

$$\begin{pmatrix} mpl_0 \\ mrs_0 \end{pmatrix} = \frac{\Lambda_p + \Lambda_w}{\Lambda_p \Lambda_w} (\rho + \delta) \begin{pmatrix} -(1-\alpha) \\ \alpha \end{pmatrix}.$$

17. Notice also that there is a t—the t at which $\dot{\omega}_t = 0$—where $\pi_t = \pi_t^w = \Pi_t^{Conflict} > 0$.

Figure 4. Price and Wage Inflation Contours for Different Degrees of Stickiness



Source: Authors' calculations.

of price adjustment λ_p increases both π and π_w but has a stronger effect on the former. The reverse holds for λ_w. For ease of illustration, we consider an economy hit by a symmetric shock $mrs_0 = -mpl_0$. This implies that when $\lambda_p = \lambda_w$, proposition 1 gives $\dot{\omega}_0 = 0$ and $\pi_0 = \pi_0^w$. In the figure, the contour levels corresponding to equal price and wage inflation meet on the 45 degree line.

Increasing either price or wage flexibility increases *both* price and wage inflation. This is the total force of the wage-price mechanism. At the same time, what happens to the real wage depends on the relative force on the two sides. Increasing λ_p tends to move us to the region below the 45 degree line, where real wages fall. Increasing λ_w has the opposite effect.

III. Demand and Supply Shocks

We now go back to the full model and trace price and wage inflation back to the general equilibrium effect of two shocks: a demand shock and a supply shock.

We show that if the economy is in an initial state that is sensitive to supply constraints, in a sense to be made precise, a positive demand shock and a negative supply shock have qualitatively similar implications on inflation. Namely, there will be a dynamic response in three phases: first, a fast increase in noncore inflation, captured here by the price of the scarce input X; then a period of sustained general inflation in prices and wages with price inflation stronger than wage inflation and real wages falling; and finally, a period of persistent wage inflation with price inflation lower than wage inflation and real wages growing back. As argued in the introduction, these dynamics seem to capture the recent post-pandemic inflationary experience well.

III.A. A Demand Shock

Consider an expansionary demand shock driven by easy monetary policy. In particular, suppose the shock is such that real spending increases to $y_0 > 0$ at date $t = 0$, and after that, it decays exponentially at rate δ, so

$$y_t = y_0 e^{-\delta t}.$$

We have not explicitly modeled monetary policy, which could be done by solving the consumers' intertemporal optimization problem and adding an interest rule to the model. However, it can be shown that the shock above translates immediately into a shock that reduces temporarily the real interest rate below its natural level (here ρ), hence stimulating consumer spending. A demand shock coming from a fiscal impulse or consumer sentiment would also have similar implications.

III.B. An Inequality for Supply-Constrained Demand Shocks

The responses of the aspirations mpl_t and mrs_t are easily derived from equations (3) and (6):

$$mpl_t = -\frac{s_X}{\epsilon} e^{-\delta t} \frac{1}{s_L} y_0 < 0, \; mrs_t = \left(\sigma s_L + \eta\right) e^{-\delta t} \frac{1}{s_L} y_0 > 0.$$

The response of the relative price of the X input (expressed in terms of labor) also follows immediately from equation (7):

$$p_{Xt} - w_t = \frac{1}{\epsilon} e^{-\delta t} n_0 > 0.$$

Given the sign of these responses, proposition 1 immediately tells us that both price and wage inflation are positive following this shock. Firms would like to pay lower real wages, given that the marginal product of labor has fallen. Consumers would like to be paid higher real wages because they are spending more and working more, so the income and substitution effects both push for a higher real marginal compensation of labor. These opposing forces produce spiral inflation, that is, conflict inflation, as discussed in the previous section.

What happens to the real wage is generally ambiguous, but proposition 1 gives us an easy condition to check and establish the sign of its response. Proposition 2 provides this condition.

PROPOSITION 2. In response to a monetary shock leading to a transitory, exponentially decaying increase in real output, price and wage inflation are both positive. Price inflation is higher than wage inflation, and consequently real wages fall at $t = 0$, if and only if the following condition is satisfied:

$$\Lambda_p \frac{s_X}{\epsilon} > \Lambda_w \left(\sigma s_L + \eta \right). \tag{16}$$

When an economy satisfies inequality (16), we say that it is supply-constrained or sensitive to supply constraints because, as we shall see, the relative scarcity of the X input driven by the ratio N_t/X_t plays a central role in price and wage inflation dynamics.

The intuition for inequality (16) is as follows.

Consider first the expression on the left-hand side, $\Lambda_p \frac{s_X}{\epsilon}$. The ratio $\frac{s_X}{\epsilon}$ captures the effect of an increase in employment on the marginal product of labor. To increase output, the economy must increase the labor input, with a fixed supply of the input X. The ratio $\frac{N_t}{X_t}$ goes up, making the X factor relatively scarcer and labor relatively abundant. How much this lowers the marginal product of labor depends on how important the input X is in the production of the final good—the share s_X—and how elastically labor can substitute for X—the elasticity ϵ. If s_X is high and ϵ is low, we get a large

effect. Finally, the coefficient Λ_p captures how quickly firms can respond to lower marginal productivity, that is, to higher marginal costs by raising nominal prices.

The expression on the right-hand side, $\Lambda_w(\sigma s_L + \eta)$, comes from the workers' side. In particular, the expression $\sigma s_L + \eta$ captures how income and substitution effects change how much workers would like to be compensated on the margin, while Λ_w captures how quickly a higher *mrs* leads to increasing nominal wages.

As we discussed in the previous section, both impulses, to *mpl* on the firms' side and to *mrs* on the workers' side, lead to mutual reactions, that is, to indirect effects: an impulse on firms' marginal costs also leads to increasing nominal wages, and an impulse on workers' marginal rate of substitution also leads to nominal price inflation. However, proposition 1 shows that the indirect effects are always weaker than the direct effects and that the presence of indirect effects does not change the relative size of the effects on the two sides. Therefore, focusing on the relative strength of the direct effects, we can safely conclude that price inflation will be higher in equilibrium than wage inflation if and only if the direct impulse on prices—the left-hand side of equation (16)—is stronger than the direct impulse on wages—the right-hand side.

III.C. An Example

Having unpacked analytically the effect of the shock at date $t = 0$, let us turn to a numerical example to look at the full dynamics and get a sense of the magnitudes involved. We focus on an example that satisfies inequality (16).

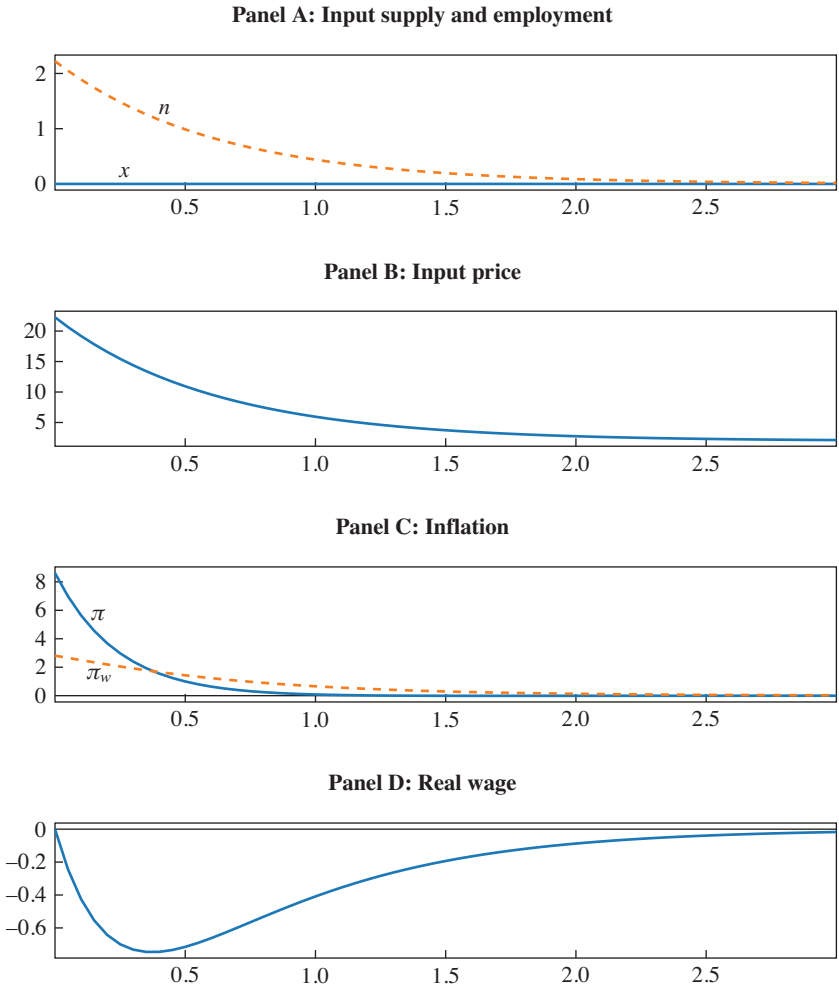
In figure 5, we plot the response to a temporary expansionary shock that increases y above potential by 2 percent on impact and converges back to potential at the rate $\delta = 1$. The parameters used are in table 1.[18]

Panel A shows the path of employment n, which is proportional to output, and the path of x, which by assumption is constant at zero. The remaining panels show the responses of different prices.

The input price is flexible, so it jumps on impact and then gradually goes back to its initial level as the shock goes away. This is shown in panel B of the figure. Notice that this panel shows the level of the input price, not its inflation rate. Inflation for that price is infinite at $t = 0$ and negative

18. All plots show log deviations from a steady state times 100 or, approximately, percentage deviations from a steady state.

Figure 5. A Supply-Constrained Demand Shock



Source: Authors' calculations.

Table 1. Parameters

Preferences	$\sigma = 1$	$\eta = 1/2$	$\rho = 0.04$
Technology	$s_X = 0.1$	$\epsilon = 0.1$	
Stickiness	$\lambda_p = 4$	$\lambda_w = 1$	

Source: Authors' calculations.

afterward. Due to perfect flexibility, P_X jumps by more than 20 percent at $t = 0$. This large increase is due to our assumption of a low elasticity of substitution between labor and the input X ($\epsilon = 0.1$), so when employment is growing too fast relative to the supply of X, the price of X reacts strongly.

The effect of the increase in the input price is to increase firms' marginal costs. The impact effect on the nominal marginal cost $w_0 - mpl_0$ is +2 percent, as the input represents 10 percent of the cost in a steady state, $s_X = 0.1$, and the elasticity is also $\epsilon = 0.1$, so the ratio $s_X/\epsilon = 1$. As we see in panel C of figure 5, this increase in marginal costs translates into fast inflation on impact: 10 percent above its steady-state level (so 12 percent inflation if we assume the central bank is keeping inflation at 2 percent in steady state).[19] This large response to a relatively small increase in marginal costs is due to our assumption of relatively flexible prices ($\lambda_p = 4$; i.e., prices reset on average every quarter), to the firms' having rational expectations and a long horizon (captured by the discount rate ρ), and, of course, to the wage response, that is, to the presence of a wage-price spiral.

On the wage side, the direct impact effect on the *mrs* is $(\sigma s_L + \eta) \times 2\% = 2.8\%$ and is close in magnitude to the effect on the marginal cost of goods, both are 2 percent. However, wages are more sticky ($\lambda_w = 1$), so the effect on wage inflation is weaker. Wage inflation is also plotted in panel C of figure 5.

The real wage falls on impact, as shown in panel D. However, as time goes by, the lower level of the real wage pushes workers to ask for nominal wage increases larger than price inflation. Wage growth eventually reverses sign and the real wage converges back to trend.

Figure 5 illustrates the three phases of adjustment mentioned in the introduction. First, very fast inflation in the sector where the supply constraints are binding, here the market for input X. Second, a phase in which price inflation is faster than wage inflation. Third, at some point wage inflation crosses price inflation and we enter the third phase in which real wages recover.

We will discuss in more depth the connection between this example and current developments at the end of this section. But first, let us look at a supply shock.

19. Notice that π_t is an instantaneous rate of inflation, expressed in annual terms. Since inflation falls relatively quickly in our example, measured quarterly inflation in the first quarter after the shock is lower than 12 percent.

III.D. A Supply Shock

Consider the same economy's response to a temporary reduction in the endowment of input X. Suppose, for now, that the central bank responds in such a way as to keep employment constant at its initial steady-state level, $n_t = 0$.

Again, the reaction of monetary policy is left implicit in the path of quantities. Since X falls, constant employment corresponds to a reduction in real output. It can be shown that this means that the central bank is increasing the real interest rate. However, as we shall see, the real rate increase that produces $n_t = 0$ is not large enough to achieve the natural allocation, given our chosen parameters.

The responses of *mpl* and *mrs* are now

$$mpl_t = \frac{s_X}{\epsilon} e^{-\delta t} x_0 < 0, mrs_t = \sigma s_X e^{-\delta t} x_0 < 0,$$

while the response of the price of good X is

$$p_{Xt} - w_t = \frac{1}{\epsilon} e^{-\delta t} n_0 > 0.$$

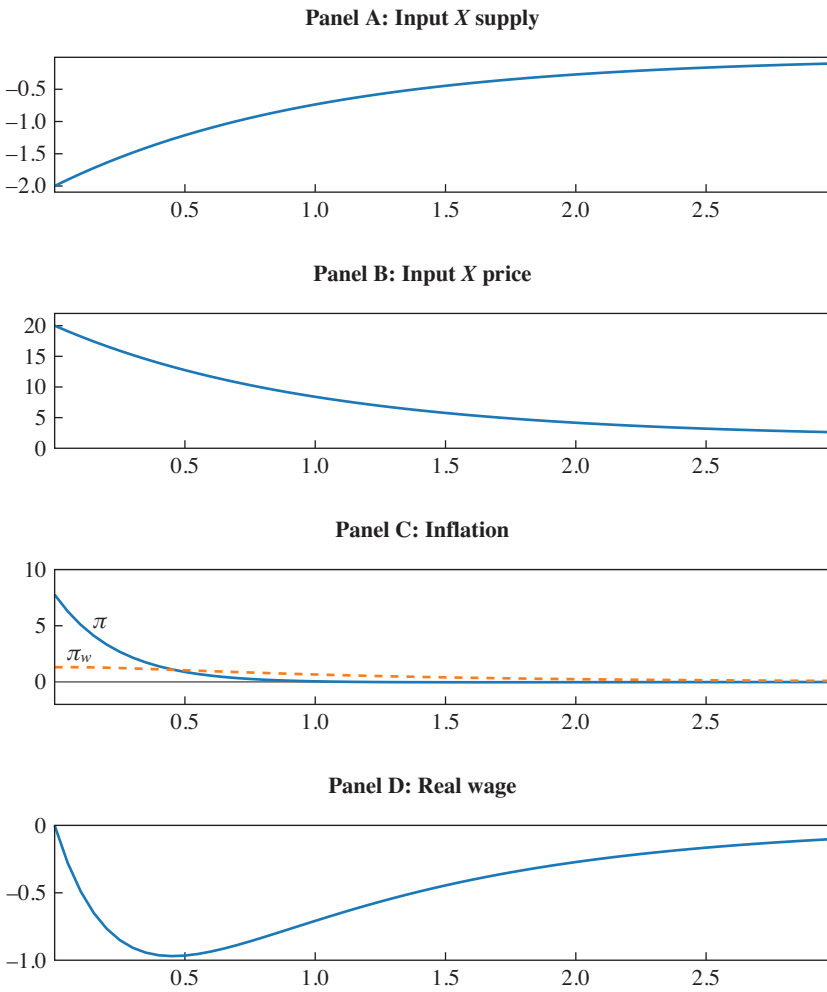
The main difference is that now the reduction in output reduces workers' *mrs* via an income effect. This weakens real wage demands. Given the parameter choices in table 1, the inflationary forces on the firms' side are still strong enough that we obtain positive wage and price inflation. In the representation in figure 3, we are in the portion of the shaded region on the left that intersects the lower left quadrant. From proposition 1, we also know that $mpl_0 < 0$ and $mrs_0 < 0$ imply that the real wage falls on impact for any parameter configuration.

The responses are illustrated in figure 6. For ease of comparison, we pick a negative shock to x_0 that produces the same increase in the input price as the positive shock to y_0 in figure 5.

While nominal wages are growing less and the real wage drop is larger than in figure 5, the overall shapes and magnitudes are not very different from the demand shock. The crucial observation here is that if we scale shocks so that the input price response is the same, we are pinning down the change in the labor-to-X ratio, as

$$p_{X0} - w_0 = \frac{1}{\epsilon}\left(n_0 - x_0\right),$$

Figure 6. A Supply Shock



Source: Authors' calculations.

and the same ratio $n_0 - x_0$ determines

$$mpl_0 = \frac{s_X}{\epsilon}\left(n_0 - x_0\right).$$

Once we choose the quantitative size of the fall in $n_0 - x_0$, we have pinned down the inflationary impulse on the firms' side.

The main difference is that in this case the wage-price spiral mechanism is weaker as workers' aspirations fall instead of increasing in the case of a supply shock. This explains why both price and wage inflation are lower in this case.

III.E. Supply Shocks and the Monetary Response

The response to the supply shock depends on how monetary policy adjusts. So far, we assumed a policy that keeps the employment path unchanged at $n_t = 0$. However, the natural level of employment depends in general on x_t. In particular, keeping employment and output at their natural levels requires that $mrs_t = mpl_t$, and n_t^* can be derived from the condition

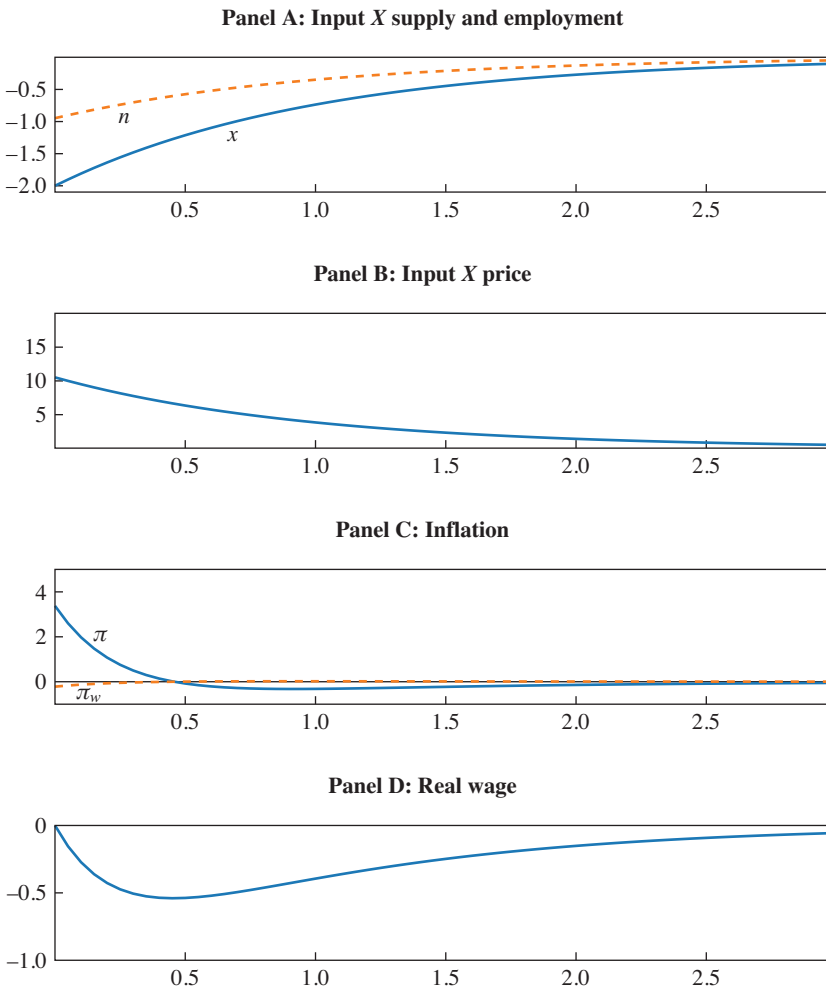
$$\sigma\left(s_L n_t^* + s_X x_t\right) + \eta n_t^* = \frac{s_X}{\epsilon}\left(x_t - n_t^*\right).$$

The responses of price and wage inflation when

$$n_t = n_t^* = \frac{\frac{1}{\epsilon} - \sigma}{\sigma s_L + \frac{s_X}{\epsilon} + \eta} s_X x_t$$

are plotted in figure 7. Since our parameterization features a low degree of substitutability between labor and the input X, we have $\frac{1}{\epsilon} - \sigma > 0$, and a reduction in x_t lowers the natural level of employment, as shown in panel A. The natural level of output $y_t^* = s_X x_t + s_L n_t^*$ is then lower for two reasons: the direct effect of a lower x_t and the lower level of natural employment. There is a clear difference in the inflation paths when quantities are at their natural levels: we see positive price inflation but negative wage inflation. This goes on as long as the real wage falls; once the real wage starts growing again, the signs of price and wage inflation flip. In other words, real wage adjustments always take place with nominal prices and wages moving in opposite directions.

Figure 7. A Supply Shock with Quantities on Their Natural Path



Source: Authors' calculations.

This is not just an outcome of our choice of parameters. When quantities are at their natural levels, we have $mrs_t = mpl_t$ and both are equal, by definition, to the natural real wage ω_t^*. The inflation equations then become

$$\pi_t = \Lambda_p \int_t^\infty e^{-\rho(s-t)}\left(\omega_s - \omega_s^*\right)ds \text{ and}$$

$$\pi_t^w = \Lambda_w \int_t^\infty e^{-\rho(s-t)}\left(\omega_s^* - \omega_s\right)ds.$$

The general result in proposition 3 follows immediately.

PROPOSITION 3. If quantities are at their natural levels, price and wage inflation π_t and π_t^w are either both zero or have opposite sign.

This result can be visualized in figure 3 by noticing that the regions where π and π^w have the same sign are either entirely above or entirely below the 45 degree line, where $mrs = mpl$.

Using the concepts introduced in section II, we can then say that if the output gap is always zero, conflict inflation is zero, that is, a wage-price spiral is not present.[20]

Behind the similar adjustment patterns illustrated in figures 5 and 6, there is a similar problem of excess demand producing positive conflict inflation. Excess demand can be caused either by a positive demand shock or a negative supply shock coupled with an insufficient monetary policy response.

However, notice also that, as is well known, an economy with both price and wage rigidities does not feature divine coincidence, so a policy of keeping the output gap at zero, that is, of keeping quantities at their flexible price levels, is not necessarily optimal in our environment. We analyze optimal policy in the next section.

Comparing figures 6 and 7 also shows that while employment falls more at the natural allocation, real wages fall less. This may seem surprising, but it is due to the fact that the dynamics of the real wage are more strongly affected by *mpl* than by *mrs*, and *mpl* is higher along the path with lower employment. A different intuition for the same phenomenon is that lower employment reduces the pressure on the market for the scarce input, as seen in panel B of both figures, weakening price inflation due to the high *X* price and increasing the real wage. Yet another intuition is that due to the

20. This result explains why conflict inflation in this model is equal to the divine-coincidence inflation of Rubbo (2020).

fact that prices of goods and nonlabor inputs are relatively more flexible than wages, the relation between real wages and employment is dominated by the labor demand side, so higher employment levels push down real wages.

III.F. Interpretation and Connections

This adjustment pattern shows both price and wage inflation, with price inflation stronger early on and wage inflation catching up later. If the central bank keeps the economy always at its flexible price allocation, this pattern will not be present, as price and wage inflation have opposite sign.

The examples presented are clearly just numerical simulations with parameters chosen mostly for clarity of exposition. Nonetheless, we believe there are some useful lessons and some interesting connections with recent experience.

DEMAND SHOCKS AND WAGE INFLATION Our model helps clarify that excess demand does not necessarily need to show up primarily through a tight labor market and high wage inflation. A commonly held view is that excessive demand works its way from a tight labor market to higher wages through the wage Phillips curve and, eventually, to higher prices. A demand shock then should produce increasing real wages. As we just showed, this is not necessarily the case. In the model, price and wage rigidities interact with general equilibrium forces on both goods and labor markets, and the direction of adjustment of the real wage is in general ambiguous. At a general level, the notion that real wages can potentially fall is obvious and commonly noted in the extreme case where nominal wages are fully rigid: in that case, the real wage must fall whenever inflation is positive.[21] Our analysis gives an easy way to interpret condition for real wages to fall or rise, clarifying the economic forces at play.

An intuitive way of making our point here is to observe that inflation is in general caused by some form of scarcity on the supply side, relative to existing demand pressures. But there are multiple inputs on the supply side, labor inputs and nonlabor inputs. Depending on the episode, scarcity can manifest itself more strongly in labor inputs or in nonlabor inputs. When nonlabor input scarcity dominates, price inflation will be faster than wage inflation.

SMALL AND LARGE ECONOMIES Many papers measure supply shocks directly in terms of changes in input prices.[22] In this paper, we emphasize the general equilibrium nature of the price shock by making the price p_X fully endogenous.

21. See, for example, figure 6.3 in Galí (2015).
22. For example, this is the strategy in the model used by Bernanke and Blanchard (2023).

It is important to remark that the degree to which p_X should be treated as endogenous or exogenous depends on the size of the economy relative to the world economy. For a small open economy that trades X frictionlessly with the rest of the world (a reasonable approximation for some energy inputs), it makes sense to redo the analysis by taking p_{Xt} as given and deriving x_t endogenously instead of shocking x_t and deriving p_{Xt} endogenously. The results for a supply shock would be similar. However, the effects of a demand shock that is completely idiosyncratic to the small open economy (that is, not correlated with a global demand shock) would be very different, as the relative scarcity of X in the world at large would not be affected by a localized shock to demand. On the other hand, a demand expansion in a large country would transmit to smaller economies as a supply shock, via the price p_X.

PASS-THROUGH FROM NONCORE TO CORE INFLATION We can identify the first phase of our three-phase responses as an initial period of high noncore inflation. Technically, the price p_X in our model does not appear directly in the Consumer Price Index (CPI), because X is only used as an input, not as a final good. Therefore, there is no distinction between core and noncore inflation in the model. However, it is easy to modify the model to allow for direct consumption of X, or for multiple sectors, some of which use X more intensively than others, and to make the distinction between core and noncore more explicit. The fact that the response of p_t lags the response of p_{Xt} shows that our model features a clear mechanism for pass-through from noncore inflation to core inflation. Recent work by Ball, Leigh, and Mishra (2022) shows empirically that this pass-through has been high in the post-pandemic period.

A related observation is that the fact that p_{Xt} is falling after jumping at $t = 0$ is not in contradiction with the fact that supply constraints are crucial for the inflation episode. It is the level of p_{Xt}, not its rate of change, that reflects the underlying scarcity in the economy, that is, a high labor to nonlabor inputs ratio $n_t - x_t$, and this scarcity is a crucial driver of the high inflation rate in goods through its effects on mpl_t.

NONLINEAR PHILLIPS CURVES Many economists have pointed out the potentially important role of a nonlinear Phillips curve in explaining recent experience.[23] Our model is linearized, but it is linearized around a steady state that captures the economy's state at the moment the shock hits. Therefore, we can easily see the effect on nonlinearities through the parameter s_X

23. See, for example, Benigno and Eggertsson (2023).

in the linearized model. That parameter is not a model's constant but depends on initial conditions. In particular, s_X is higher if the initial steady state features a relatively high initial ratio N_t/X_t. In other words, if the X input is already relatively scarce when the shock hits, the effects of the shock on inflation will be magnified. It would be interesting to explore model extensions in which the elasticity ϵ is also endogenous and depends on the state of the economy.

Notice that the nonlinearity we are pointing out here is not nonlinearity in the wage Phillips curve, which is the one that has received more attention, but rather nonlinearity in the response of nonlabor input prices, which affects the price Phillips curve.[24]

PROFITS A possible interpretation of the scarce input X is not as a market-supplied input but rather as capturing fixed production capacity and other bottlenecks at the firm level. The formal analysis is slightly different when the input is fixed at the firm level instead of being fixed economy-wide and frictionlessly traded.[25] But the qualitative responses are similar.

There is, however, a marked difference in interpretation between a model with a market-supplied input X and a model with fixed capacity. In the first model, observed profit margins at the firm level fall in response to the shocks analyzed because nominal prices increase less than marginal costs due to stickiness. In the second model, observed profit margins increase because firm profits include the shadow price of the scarce input X, which increases sharply in all our examples.

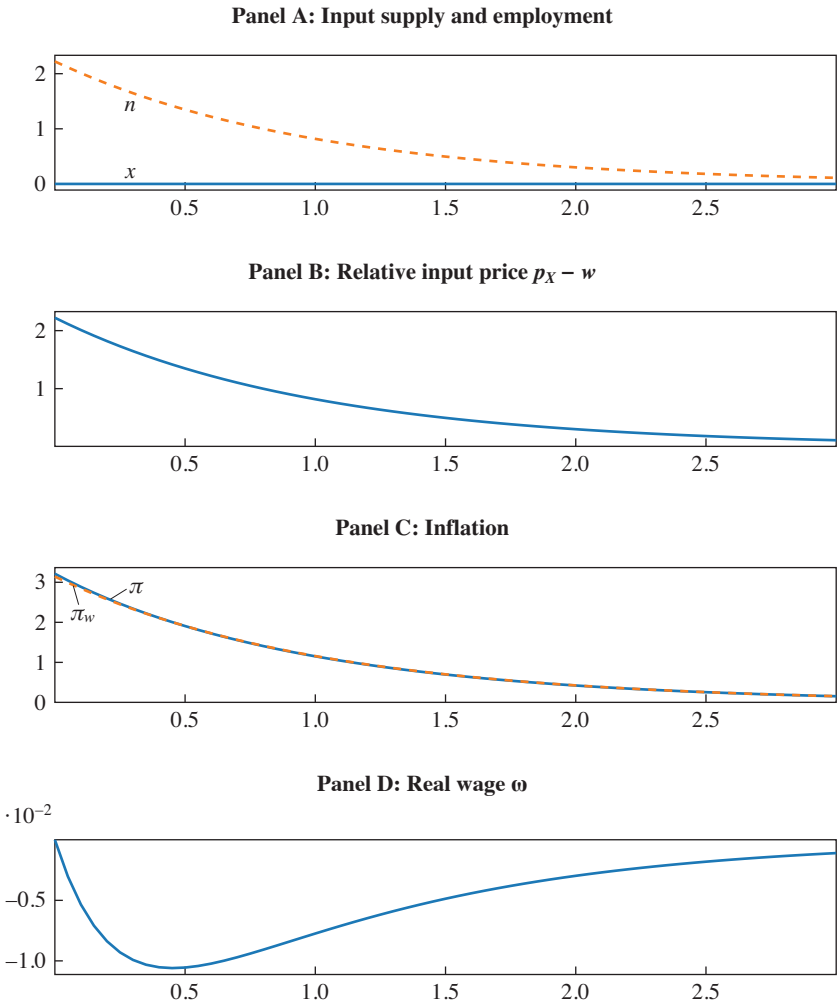
THE ROLE OF ϵ In our examples, we have used a low elasticity $\epsilon = 0.1$. This low elasticity plays two roles: it magnifies the response of p_{Xt}, explaining the initial jump in noncore inflation, and it magnifies the response of *mplt*, explaining the prolonged inflation episode. To see the central role of this parameter, consider an example with all the same assumptions of our demand shock in figure 5 but assume a Cobb-Douglas production function, with $\epsilon = 1$. The responses are plotted in figure 8.

Two differences stand out compared to our baseline parametrization. First, there is a smaller response of the relative price of the X input in panel B. With higher elasticity, the relative scarcity of X has a smaller price effect (the effect is proportional to $1/\epsilon$, so it falls by a factor of 10).

24. Comin, Johnson, and Jones (2023) use occasional binding constraints to study a model with a similar nonlinearity in the price Phillips curve.

25. In particular, a model with firm-specific, non-traded X is a model with decreasing returns to labor at the firm level, which produces strategic complementarity in pricing that is absent in our model with constant returns.

Figure 8. A Demand Shock with Higher Elasticity of Substitution



Source: Authors' calculations.

This implies a smaller overall inflation response. Second, the responses of wage and price inflation are almost indistinguishable, and consequently, the real wage is not affected. This is because the response of *mpl* is weaker while the response of *mrs* is unchanged (as we keep the value of *sL* unchanged in the two examples).

This suggests that, at the aggregate level, to capture episodes in which the relative scarcity of nonlabor inputs triggers an inflationary episode, with a lagged response of wage inflation, a low degree of elasticity at the aggregate level is a needed ingredient.

IV. Optimal Policy

In the previous section, we looked at economies in which the central bank unnecessarily stimulates the economy (demand shock) or the central bank responds weakly to a supply shock, so as to allow for both price and wage inflation (the supply shock with $n_t = 0$). The first example is a policy mistake by construction. Of course, due to imperfect information and lags in the effects of monetary policy, mistakes can happen. However, in this section, we focus on the second shock, a supply shock, and ask what the optimal response is. Throughout, we assume monetary policy has perfect information on the underlying shocks and instantaneous control on the level of real activity.

The questions we address in this section are two. Is it possible that, following a supply shock, the optimal response is to let the economy overheat, that is, to choose a positive output gap $y_t - y_t^* > 0$? Is it possible that the optimal response entails both positive price and wage inflation?

It is well known that divine coincidence fails in our environment. But that is just a statement about feasibility: an outcome with no inflationary distortions, $\pi_t = \pi_t^w = 0$, and a zero output gap, $y_t = y_t^*$, are not feasible in our model. The real wage needs to move in the flexible price equilibrium and that is incompatible with zero nominal inflation in p_t and w_t. Our contribution here is to characterize the signs of the deviations of π_t, π_t^w, and $y_t - y_t^*$ from zero under optimal policy.

In particular, proposition 3 above tells us that if the central bank chooses $y_t = y_t^*$, then the signs of π_t and π_t^w will always be opposite. In other words, with a zero output gap, the adjustment in the real wage never requires *both* price and wage inflation. Therefore, one could conjecture that generalized inflation, that is, inflation in both prices and wages, is never optimal. However, a zero output gap is not necessarily optimal, so that conjecture is not generally correct.

IV.A. Optimal Policy Problem

Following standard steps, the objective function of the central bank can be derived as a quadratic approximation to the social welfare function:

$$\int_0^\infty e^{-\rho t}\frac{1}{2}\left[-\left(y_t - y_t^*\right)^2 - \Phi_p \pi_t^2 - \Phi_w\left(\pi_t^w\right)^2\right]dt. \tag{17}$$

Deviations from first-best welfare come from two distortions: output deviations from its natural level, that is, from the level that equalizes the marginal benefit of producing goods with its marginal cost in terms of labor effort; and inflation in prices and wages that causes inefficient dispersion in relative prices of different varieties. The terms in equation (17) reflect these distortions. The values of the coefficients Φ_p and Φ_w depend on the model parameters and are derived in the online appendix.

The natural level of the real wage following a supply shock is

$$\omega_t^* = \frac{s_X}{\epsilon}\frac{\sigma+\eta}{\sigma s_L + \frac{s_X}{\epsilon} + \eta}x_t.$$

We can then express *mpl* and *mrs* in terms of the natural real wage and deviations of employment from its natural path

$$mpl_t = \omega_t^* - \frac{s_X}{\epsilon}\left(n_t - n_t^*\right) \text{ and} \tag{18}$$

$$mrs_t = \omega_t^* + \left(\sigma s_L + \eta\right)\left(n_t - n_t^*\right). \tag{19}$$

The optimal policy problem is to maximize equation (17), subject to the constraints coming from the price-setting (10) and (11), the real wage dynamic equation

$$\dot{\omega}_t = \pi_t^w - \pi_t$$

and the aggregate production function expressed as

$$y_t = s_L n_t + s_X x_t.$$

The optimality conditions that characterize an optimal policy are derived in the online appendix.

IV.B. Examples

We now consider examples that illustrate a variety of possible outcomes.

It helps the interpretation of the policy trade-offs to focus on the simple case of a permanent shock to x_t. With this shock, in all our examples, in the long run, the real wage is permanently lower and so are *mpl* and *mrs*, so that the economy eventually reaches a new steady state with zero inflation and zero output gap. To reach that new steady state requires ω_t to fall. This can be achieved by many combinations of price and wage inflation or deflation, as long as price inflation is larger than wage inflation. The question is, what is the optimal way to get there?

EXAMPLE 1. A SYMMETRIC CASE Our first example is an economy with parameters that have the following properties: the welfare costs of wage and price inflation enter symmetrically the objective function, $\Phi_p = \Phi_w$; wages and prices are equally sticky, $\Lambda_p = \Lambda_w$; and the output gap has symmetric effects on *mpl* and *mrs*.[26]

Figure 9 illustrates optimal policy outcomes in this example. Given the symmetry of the problem, the reduction in real wages is achieved by spreading the adjustment equally between nominal wage deflation and nominal price inflation. The output gap is kept exactly at zero. This example is clearly a knife-edge case and relies on the symmetry of the parameters. As soon as we abandon this symmetry things get more interesting.

EXAMPLE 2. A HOT ECONOMY In the second example, the parameters chosen imply that the welfare cost of wage inflation is larger than that of price inflation, $\Phi_p < \Phi_w$, and wages are more sticky than prices, $\Lambda_p > \Lambda_w$.[27] We still have a set of parameters that implies roughly symmetric effects of the output gap on *mpl* and *mrs*, but the differences are sufficient to obtain a quite different result. Figure 10 illustrates optimal policy outcomes in

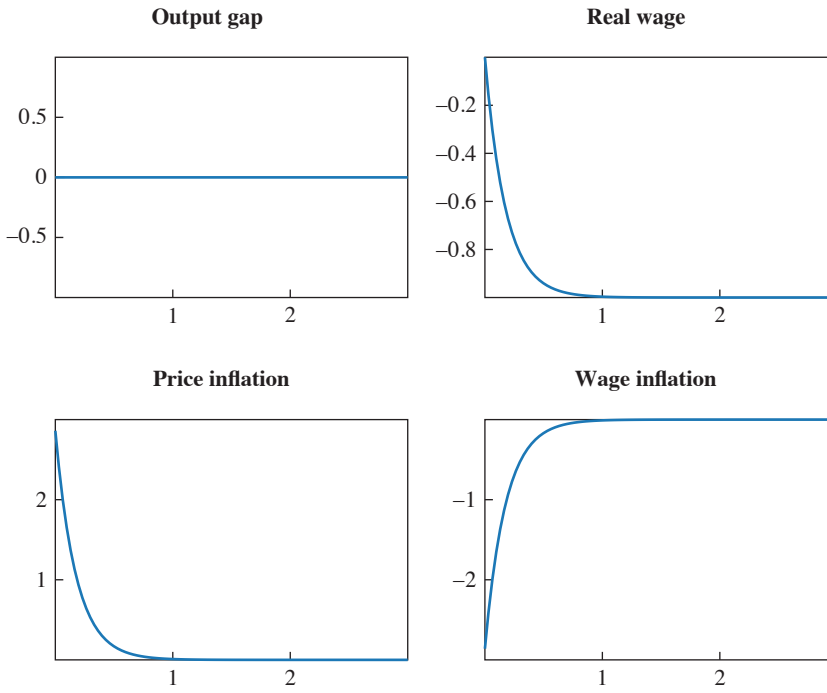
26. The following parameters satisfy these conditions and are used in the numerical example:

$\sigma = 1$	$\eta = 0$	$\rho = 0.05$	
$s_X = 1/2$	$\epsilon = 1$	$\varepsilon_C = 1.5$	$\varepsilon_L = 3$
$\lambda_p = 4$	$\lambda_w = 4$		

27. The parameters are as follows:

$\sigma = 1$	$\eta = 0$	$\rho = 0.05$	
$s_X = 0.1$	$\epsilon = 1$	$\varepsilon_C = 1.5$	$\varepsilon_L = 4$
$\lambda_p = 4$	$\lambda_w = 2$		

Figure 9. A Symmetric Example



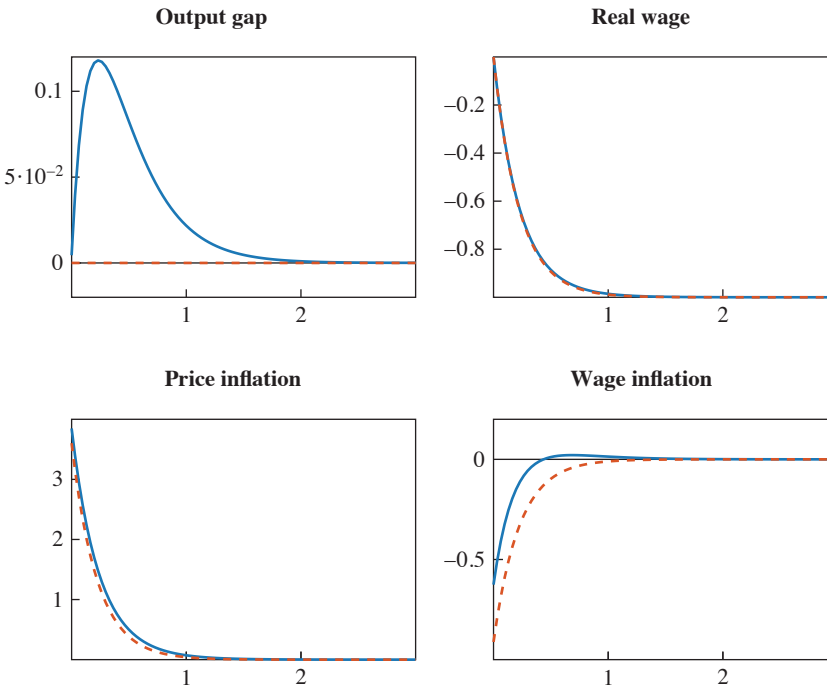
Source: Authors' calculations.

this case. For comparison, in the figure we also plot outcomes under a zero output gap policy (dashed lines).

In this second example, it is optimal to have a positive output gap throughout the transition. Recall from equations (10)–(11) and (18)–(19) that increasing the output gap has two direct effects: by decreasing *mpl*, it leads to higher price inflation; and by increasing *mrs*, it leads to higher wage inflation. If we start at a zero output gap policy with positive price inflation and negative wage inflation, the effect can be welfare improving because the welfare cost of price inflation is smaller than the welfare cost of wage deflation.

The role of $\Lambda_p > \Lambda_w$ is subtler and has to do with dynamics. With $\Lambda_p > \Lambda_w$, a higher output gap also implies a faster declining real wage. Since a lower real wage in the future requires less adjustment, lowering the real wage today is welfare improving from a dynamic point of view. Therefore, a parameterization with $\Lambda_p > \Lambda_w$ makes it easier to obtain examples with a welfare-improving positive output gap.

Figure 10. An Optimal Hot Economy



Source: Authors' calculations.

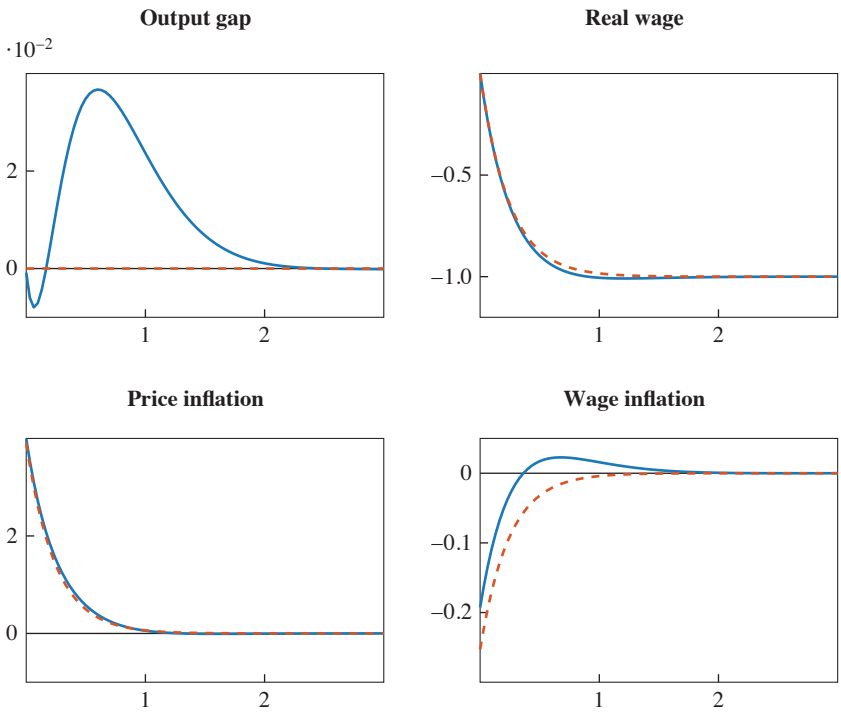
Notice that it is also possible to choose parameters that imply that the welfare costs of price inflation are relatively larger than those of wage inflation, and to obtain examples in which it is optimal to run a negative output gap in the transition.

EXAMPLE 3. GENERALIZED INFLATION AND A HOT ECONOMY Our third example is a variant on the second example, with an even larger welfare cost associated to wage dispersion (a larger Φ_w), a larger distance between price and wage stickiness, and a smaller value of the elasticity of substitution between labor and the X input, ϵ, which implies that running a hot economy has larger benefits in terms of lowering the real wage by having a larger effect on firms' marginal costs and thus on price inflation.[28]

28. The parameters are as follows:

$\sigma = 1$	$\eta = 0$	$\rho = 0.05$	
$s_X = 0.1$	$\epsilon = 0.1$	$\varepsilon_C = 1.5$	$\varepsilon_L = 8$
$\lambda_p = 4$	$\lambda_w = 1$		

Figure 11. An Example with Generalized Inflation and a Hot Economy



Source: Authors' calculations.

The parametric choices above amplify the forces we saw in example 2, and they imply that there is an interval during the transition in which the optimal policy yields both a hot economy ($y_t > y_t^*$) and generalized price and wage inflation ($\pi_t > 0$ and $\pi_t^w > 0$).[29]

This result is surprising from a static point of view (see figure 11). Given the welfare function (17), at any point in time in which $y_t > y_t^*$, $\pi_t > 0$, and $\pi_t^w > 0$, it is welfare improving, from a static point of view, to reduce y_t, as it unambiguously lowers π_t and π_t^w and leads to an increase in the current payoff. However, from a dynamic perspective, there is an additional argument. Increasing y_t at time t has the effect of increasing π_s and π_s^w in all previous periods due to the forward-looking element in price setting. This entails welfare gains in early periods in the transition in which

29. Notice that these qualitative features can actually be seen in example 2 too, but it is useful to choose an example where they are more clearly visible.

$\pi_s^w < 0$. Through this forward-looking force, a positive output gap later in the transition can be beneficial even if, at that point, $\pi_t^w > 0$.

Now, while this example is theoretically interesting, it does have the flavor of an overly sophisticated form of forward guidance. Therefore, we do not think it provides a strong argument in favor of policies that deliver $y_t > y_t^*$, $\pi_t > 0$, and $\pi_t^w > 0$ at the same time. In the context of the present model, given the distortions it captures, it is hard to make a compelling practical case that the combination of a hot economy with positive wage and price inflation is a desirable outcome, even in response to a supply shock and even in the presence of inelastic supply constraints.[30]

V. Adaptive Expectations and Real Rigidities

The model with rational expectations analyzed so far has two embedded features: the effect of any shock tends to be front-loaded, as agents perfectly anticipate its future effects on prices; and there is no room for persistent deviations of inflation expectations from target, as agents anticipate the economy will go back to its initial steady state. We now explore variants of the model that deviate from rational expectations and allow for more inertial responses by introducing two ingredients: adaptive expectations on expected inflation and a gradual adjustment of price setters' and wage setters' relative price objectives. For this second ingredient we use the label "real rigidities."

The objective of this section is twofold. First, by allowing for inertial responses, we allow the feedback between prices and wages to play out more explicitly over time: shocks that produce high prices in the goods market only gradually lead to higher wage demands in the labor market. In other words, the wage-price spiral, instead of playing out in the "virtual time" of best responses, plays out in the observed dynamics of prices and wages. Second, by allowing for deviations of inflation expectations from target, we capture the common concern of central bankers that prolonged episodes of high inflation may lead to de-anchoring of inflation expectations.

From an empirical perspective, we show that adaptive expectations and inertia reinforce the main prediction of the baseline model in section III: there is a lagged and persistent increase in wage inflation following a large increase in price inflation. However, the medium-term implications are

30. This does not mean that such a case could not maybe be made in richer models, which capture, just to make an example, the benefits of labor reallocation as in Guerrieri and others (2021). But that is clearly outside the scope of this paper.

different depending on the sources of inertia: if inertia is mostly due to de-anchoring, inflation can take a long time to go back to target, absent a recession; if instead inertia is mostly due to real rigidities, then a path of immaculate disinflation is possible.

Let us begin by rewriting the price-setting conditions making explicit agents' expectations. Letting E_t^f and E_t^w denote firms' and workers' expectations, we can write

$$p_t^* = \left(\rho + \lambda_p\right) E_t^f \int_t^\infty e^{-(\rho+\lambda_p)(\tau-t)} \left(w_\tau + s_X\left(p_{X\tau} - w_\tau\right)\right) d\tau$$

$$= w_t + \left(\rho + \lambda_p\right) E_t^f \int_t^\infty e^{-(\rho+\lambda_p)(\tau-t)} s_X\left(p_{X\tau} - w_\tau\right) d\tau$$

$$+ E_t^f \int_t^\infty e^{-(\rho+\lambda_p)(\tau-t)} \dot{w}_\tau d\tau.$$

Reset prices are decomposed in three components: the current nominal wage, the expected path of the relative price of input X versus labor, and the expected path of future wage inflation.

We assume that agents expect a constant inflation rate over the future horizon

$$E_t^f \dot{w}_t = \pi_t^{w,e},$$

and expected inflation is driven by the simple adaptive, constant-gain rule

$$\dot{\pi}_t^{w,e} = \gamma\left(\dot{w}_t - \pi_t^{w,e}\right). \tag{20}$$

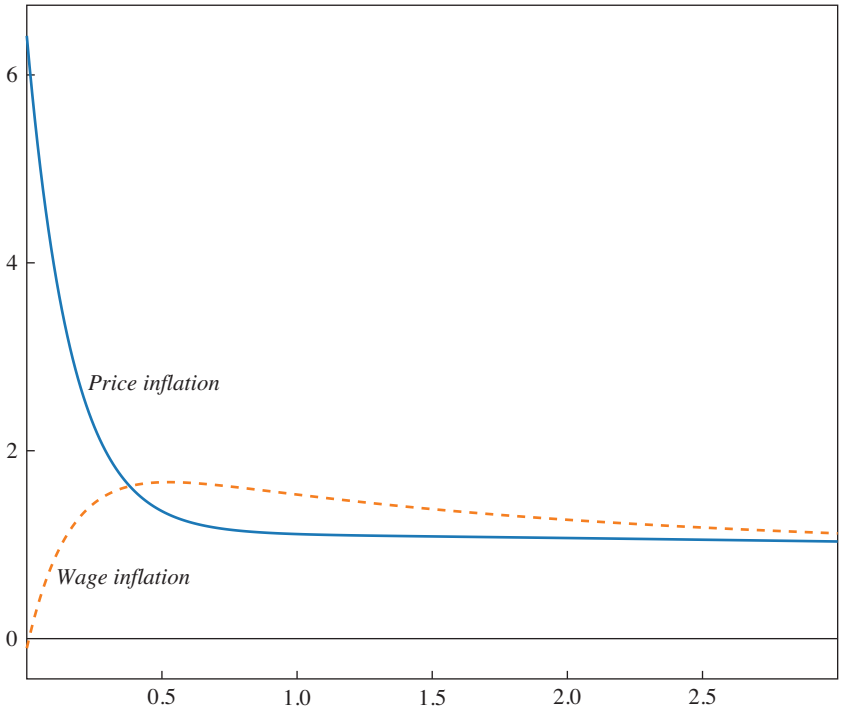
Moreover, we assume that agents perfectly anticipate the path of real variables n_t, x_t, and y_t, and can deduce the path of the relative price $p_{Xt} - w_t$ from the equilibrium condition in factor markets

$$x_t - n_t = -\epsilon\left(p_{Xt} - w_t\right).$$

Combining these assumptions with exponentially decaying, one-time shocks at date zero, as in section III, we can substitute in the expression above for p_t^*, substitute in the inflation equation (8), and obtain the following:

$$\dot{p}_t = \lambda_p \left[\frac{s_X}{\epsilon} \frac{\rho + \lambda_p}{\rho + \lambda_p + \delta}\left(n_t - x_t\right) - \left(p_t - w_t\right)\right] + \frac{\lambda_p}{\rho + \lambda_p} \pi_t^{w,e}. \tag{21}$$

Figure 12. A Supply Shock with Adaptive Expectations



Source: Authors' calculations.

Similar steps on the wage-setting side of the model lead to

$$\dot{w}_t = \lambda_w \left[\frac{\rho + \lambda_w}{\rho + \lambda_w + \delta} \left(\sigma y_t + \eta n_t \right) - \left(w_t - p_t \right) \right] + \frac{\lambda_w}{\rho + \lambda_w} \pi_t^e, \tag{22}$$

where price inflation follows the adaptive rule

$$\dot{\pi}_t^e = \gamma \left(\dot{p}_t - \pi_t^e \right). \tag{23}$$

Equations (20)–(23) can be solved forward for any given initial condition w_0, p_0.

AN EXAMPLE OF DE-ANCHORING Figure 12 shows the response of inflation to a supply shock in a numerical example analogous to the one shown in figure 6, except for the assumption of adaptive expectations. The parameters

are the same as in table 1, and we set $\gamma = 1$. There are two main differences from the case of rational expectations. First, wage inflation is weaker on impact and only picks up gradually, as initially workers do not anticipate higher prices and so do not start trying to catch up until their purchasing power has actually been eroded by past inflation.[31] Second, there is a very persistent effect on inflation, due to the learning dynamics. Since ρ is small, the coefficients on the expected inflation terms on the right-hand side of equations (21)–(22) are close to one. This implies that even though all quantities and all relative price targets for workers and firms have gone back to steady state, we can have a prolonged period of self-sustaining inflation. This is a case of de-anchoring in which the only way to go back to target inflation faster is for the central bank to keep activity low for some time.

The wage-price spiral is active in the self-sustaining phase of prolonged inflation, but it is exactly balanced on the two sides, so real wages remain constant.

AN EXAMPLE WITH REAL RIGIDITIES We now consider a different source of inertia, due to a gradual adjustment of the relative price targets of price and wage setters. In particular, we assume that changes in real marginal costs and the marginal rate of substitution between consumption and leisure only gradually change the behavior of price and wage setters. We replace the inflation dynamics above, equations (21)–(22), with the following equations:

$$\dot{p}_t = \lambda_p \left[a_t^p - \left(p_t - w_t \right) \right] + \frac{\lambda_p}{\rho + \lambda_p} \pi_t^{w,e} \text{ and}$$

$$\dot{w}_t = \lambda_w \left[a_t^w - \left(w_t - p_t \right) \right] + \frac{\lambda_w}{\rho + \lambda_w} \pi_t^e.$$

The real aspirations of price setters and wage setters, a_t^p and a_t^w, follow the adjustment equations

$$\dot{a}_t^p = \xi_p \left[\frac{s_X}{\epsilon} \frac{\rho + \lambda_p}{\rho + \lambda_p + \delta} \left(n_t - x_t \right) - a_t^p \right] \text{and}$$

31. Notice that given that n is kept on its pre-shock path ($n = 0$) and that output falls due to the supply shock ($y_0 = s_x x_0 < 0$), there is an income effect that depresses the real wage demands of workers on impact, causing a very small initial nominal wage deflation, which is barely visible in the figure.

$$\dot{a}_t^w = \xi_w \left[\frac{\rho + \lambda_w}{\rho + \lambda_w + \delta} \left(\sigma y_t + \eta n_t \right) - a_t^w \right].$$

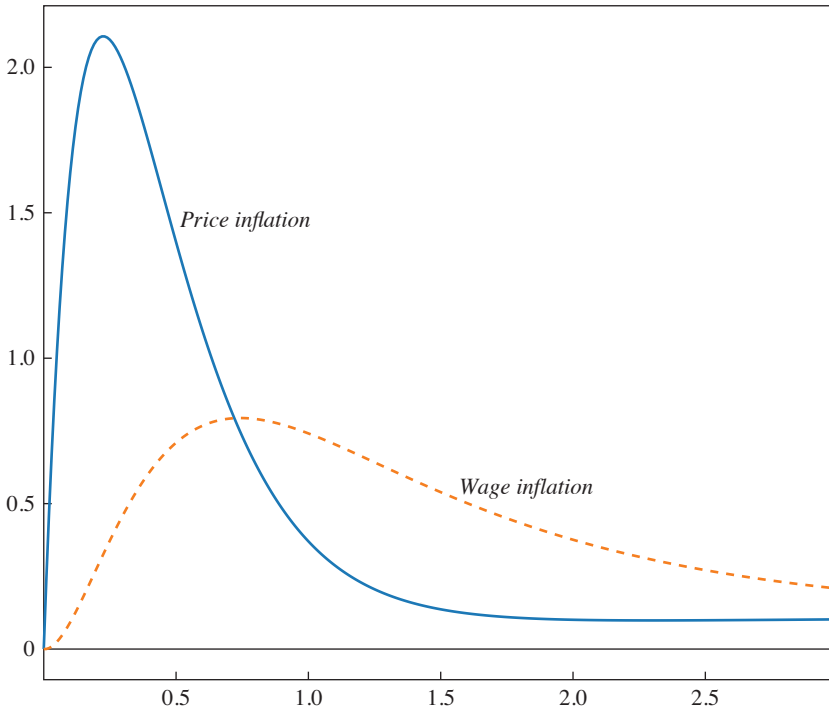
Aspirations are driven by the same forces that drive them in the baseline model, which in the case of firms are anticipated real input prices captured by the term $\frac{s_X}{\epsilon} \frac{\rho + \lambda_p}{\rho + \lambda_p + \delta} \left(n_t - x_t \right)$, and in the case of workers are anticipated marginal rates of substitution between consumption and leisure captured by $\frac{\rho + \lambda_w}{\rho + \lambda_w + \delta} \left(\sigma y_t + \eta n_t \right)$. However, these forces only gradually modify the aspirations of firms in terms of the desired margins ($p_t - w_t$ for the firms and $w_t - p_t$ for the workers).

We assume that the inflation expectations $\pi_t^{w,e}$ and π_t^e still follow the learning processes equations (20) and (23), so this version of the model includes both inertia caused by slow adjustment of inflation expectations and inertia caused by real rigidities. The choice to combine the two is because an interpretation of the real rigidities here is also some form of bounded rationality in processing observed changes in input prices and changes in labor market conditions, and combining that with perfect foresight on future price paths seems less natural. However, to focus on the role of real rigidities, we choose a parameterization with a lower $\gamma = 0.1$, relative to the parameterization used for figure 12, so inflation expectations play a more limited role. For the parameters ξ_p and ξ_w, we experiment with values equal to four and one, so the degree of real rigidity in the goods and labor market mirror the degree of nominal rigidity (captured by λ_p and λ_w). The inflation responses to the same supply shock used above are reported in figure 13.

In this economy, both price and wage inflation display hump-shaped responses, and the wage response is more delayed and more persistent than in the rational expectations baseline. The delay in the wage response is essentially due to the same reason as in the model with only adaptive inflation expectations: wage setters only start to demand higher nominal wages when price inflation has been going on for a while and has moved real wages away from their aspirations. The additional delay here is because of the fact that prices also take longer to respond due to the real rigidity in price setting.[32]

32. The real rigidity in wage setting does not really play an important role in this simulation because with a pure supply shock to x, the effect on $\sigma y + \eta n$ is very small, so workers' aspirations are essentially constant at zero. In line with this observation, simulations with larger and smaller values of ξ_w produce responses very similar to those in figure 13. Of course, in the case of other shocks this is no longer the case.

Figure 13. A Supply Shock with Adaptive Expectations



Source: Authors' calculations.

The example in figure 13 comes closest to capturing an immaculate inflation-disinflation scenario. The shock causes persistent responses of prices and wages. The persistence is purely due to the fact that price setters take some time to respond and wage inflation follows with further delay because wage setters only start responding after price setters have increased the price level enough to lower $w - p$. The persistence of wage inflation in this scenario is not a symptom of persistent overheating in the labor market but of a gradual return to pre-shock trends for the real wage.

VI. Conclusion

We explored the wage-price spiral in a canonical model of price and wage setting.

Interpreting inflation as the outcome of inconsistent aspirations for the real wage (or other relative prices) opens the door to many theoretical and

empirical questions. We are especially interested in extending our work to explore potential sources of inertia in the inflation process, expanding the models explored in section V.

In the model analyzed here there is an instantaneous connection between the output gap and the real wage aspirations of workers and firms. However, it is plausible that workers' real wage aspirations respond gradually to changes in labor market conditions. Similarly, changes in goods market conditions could slowly affect firms' expected profit margins. These are sources of inertia in inflation that come from agents' views on relative prices and so are different from sources of inertia tied to future inflation expectations, which most research has focused on. Even if inflation expectations are well anchored, it is possible for inflation to persist if the disagreement between firms and workers is inertial. On the empirical front, while there is a lot of literature measuring inflation expectations, there has been limited effort so far at measuring workers' and firms' aspirations for real pay and real profit margins.

ACKNOWLEDGMENTS We thank Olivier Blanchard, Jason Cummins, Jordi Galí, Bob Rowthorn, Ayşegül Şahin, and Jón Steinsson for comments and suggestions on a previous draft of this paper. Pedro Bruera and Valeria Vazquez provided excellent research assistance.

References

Amiti, Mary, Sebastian Heise, Fatih Karahan, and Ayşegül Şahin. 2023. "Inflation Strikes Back: The Role of Import Competition and the Labor Market." *NBER Macroeconomics Annual* 38.

Ball, Laurence, Daniel Leigh, and Prachi Mishra. 2022. "Understanding US Inflation during the COVID-19 Era." *Brookings Papers on Economic Activity*, Fall, 1–54.

Barlevy, Gadi, and Luojia Hu. 2023. "Unit Labor Costs and Inflation in the Non-Housing Service Sector." Chicago Fed Letter 477. https://www.chicagofed.org/publications/chicago-fed-letter/2023/477.

Benigno, Pierpaolo, and Gauti B. Eggertsson. 2023. "It's Baaack: The Surge in Inflation in the 2020s and the Return of the Non-Linear Phillips Curve." Working Paper 31197. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31197.

Bernanke, Ben S., and Olivier J. Blanchard. 2023. "What Caused the U.S. Pandemic-Era Inflation?" Working Paper 86. Washington: Hutchins Center on Fiscal and Monetary Policy at Brookings. https://www.brookings.edu/wp-content/uploads/2023/06/WP86-Bernanke-Blanchard_6.13.pdf.

Blanchard, Olivier J. 1986. "The Wage Price Spiral." *Quarterly Journal of Economics* 101, no. 3: 543–66.

Blanchard, Olivier J. 2021. "In Defense of Concerns over the $1.9 Trillion Relief Plan." Washington: Peterson Institute for International Economics.

Blanchard, Olivier J., and Jordi Galí. 2007a. "The Macroeconomic Effects of Oil Shocks: Why Are the 2000s So Different from the 1970s?" Working Paper 13368. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w13368/w13368.pdf.

Blanchard, Olivier J., and Jordi Galí. 2007b. "Real Wage Rigidities and the New Keynesian Model." *Journal of Money, Credit and Banking* 39, no. s1: 35–65.

Bruno, Michael, and Jeffrey Sachs. 1985. *Economics of Worldwide Stagflation*. Cambridge Mass.: Harvard University Press.

Christiano, Lawrence J., and Martin Eichenbaum. 1992. "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations." *American Economic Review* 82, no. 3: 430–50.

Comin, Diego A., Robert C. Johnson, and Callum J. Jones. 2023. "Supply Chain Constraints and Inflation." Working Paper 31179. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31179.

Erceg, Christopher J., Dale W. Henderson, and Andrew T. Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46, no. 2: 281–313.

Gagliardone, Luca, and Mark Gertler. 2023. "Oil Prices, Monetary Policy and Inflation Surges." Working Paper 31263. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31263.

Galí, Jordi. 2015. *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications*, 2nd edition. Princeton, N.J.: Princeton University Press.

Giannoni, Marc P., and Michael Woodford. 2005. "Optimal Inflation-Targeting Rules." In *The Inflation-Targeting Debate*, edited by Ben S. Bernanke and Michael Woodford. Chicago: University of Chicago Press.

Guerrieri, Veronica, Guido Lorenzoni, Ludwig Straub, and Iván Werning. 2021. "Monetary Policy in Times of Structural Reallocation." In *Economic Policy Symposium Proceedings: Macroeconomic Policy in an Uneven Economy*. Jackson Hole, Wyo.: Federal Reserve Bank of Kansas City.

Kabaca, Serdar, and Kerem Tuzcuoglu. 2023. "Supply Drivers of US Inflation since the COVID-19 Pandemic." Working Paper 2023-19. Ottawa: Bank of Canada. https://www.bankofcanada.ca/2023/03/staff-working-paper-2023-19/.

Lorenzoni, Guido, and Iván Werning. 2022. "Inflation Is Conflict." Working Paper. https://economics.mit.edu/sites/default/files/inline-files/conflict%20inflation_0.pdf.

Rotemberg, Julio J., and Michael Woodford. 1992. "Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity." *Journal of Political Economy* 100, no. 6: 1153–207.

Rowthorn, Robert E. 1977. "Conflict, Inflation and Money." *Cambridge Journal of Economics* 1, no. 3: 215–39.

Rubbo, Elisa. 2020. "Networks, Phillips Curves, and Monetary Policy." Working Paper. https://scholar.harvard.edu/files/elisarubbo/files/rubbo_jmp.pdf.

Comments and Discussion

COMMENT BY

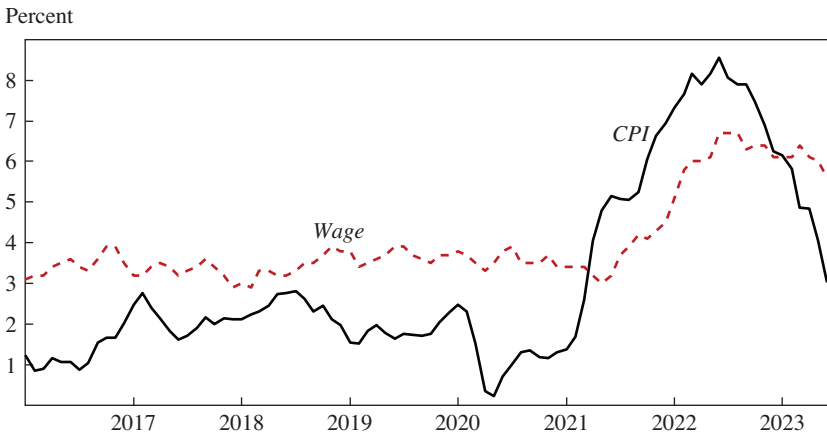
JORDI GALÍ Lorenzoni and Werning deal with a subject that is central to macroeconomics: the sources and mechanisms behind inflation fluctuations. Interest in that subject has only been enhanced by the recent high inflation episode. More specifically, they revisit the potential role of wage-price spirals as a factor of inflation persistence using a New Keynesian model with staggered price and wage setting à la Erceg, Henderson, and Levin (2000) as a reference framework. Their analysis yields a number of interesting results, including a connection between wage-price spirals and the concept of "conflict inflation," which they introduced in earlier work (Lorenzoni and Werning 2022). The paper contains many insights, of which I will single out the discussion of the potential role of two departures from the standard model as sources of inflation persistence, namely, the introduction of expectations de-anchoring and real rigidities.

My discussion is organized as follows. Firstly, I raise a caveat regarding the authors' characterization of the recent wage and price developments that motivate the paper. I then contrast the notion of inflation as conflict proposed in the paper with a more conventional interpretation of wage spirals. Next, I will discuss the connection between wage-price spirals and conflict inflation and relate some of the paper's normative findings to the existing literature. Finally, I will discuss the extensions of the model incorporating adaptive expectations and real rigidities.

RECENT WAGE AND PRICE DEVELOPMENTS REVISITED While the focus of Lorenzoni and Werning's paper is theoretical, its motivation is driven by the wage and price developments observed in the wake of the COVID-19

Brookings Papers on Economic Activity, Fall 2023: 368–393

Figure 1. CPI and Wage Inflation

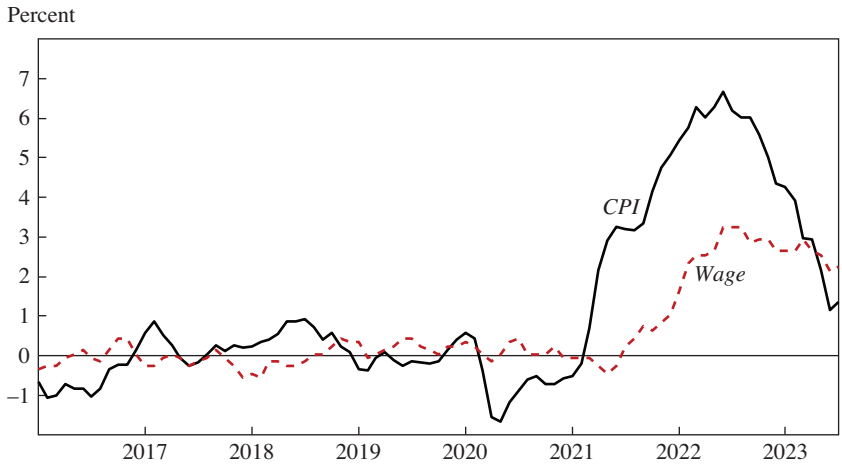


Source: CPI data from Bureau of Labor Statistics, retrieved from FRED; and wage inflation data from Wage Growth Tracker, Atlanta Fed.

pandemic and the war in Ukraine. Figure 1 summarizes these developments by displaying year-on-year US price and wage inflation from 2016 onward, using the Consumer Price Index (CPI) and the Federal Reserve Bank of Atlanta's wage index, respectively. The figure reveals the temporal pattern stressed by the authors, with wage inflation lagging price inflation both on the way up—with the real wage declining as a result—and on the way down—with wage inflation remaining roughly unchanged over the past year even in the face of a marked decline in price inflation—with the consequent increase in the real wage. That observation had led, in the authors' words, to "the concern . . . that higher wage growth would prevent inflation from going back to target, or even set off an out-of-control wage-price spiral." A central message of the paper is that such a concern is likely to be unwarranted, for the observed pattern is precisely the one that a standard model, calibrated in a way consistent with the evidence on the relative stickiness of prices and wages, would predict in response to either an expansionary demand shock or an adverse supply shock (both persistent, but not permanent) in an environment in which the monetary policy rule guarantees the return of price inflation to its intended target.

Here I would like to point out a caveat in the authors' analysis: the fact that price inflation and wage inflation display different underlying trends may distort the interpretation of figure 1 and its connection with the subsequent model simulations (which abstract from those differential trends). More

Figure 2. CPI and Wage Inflation (Demeaned)

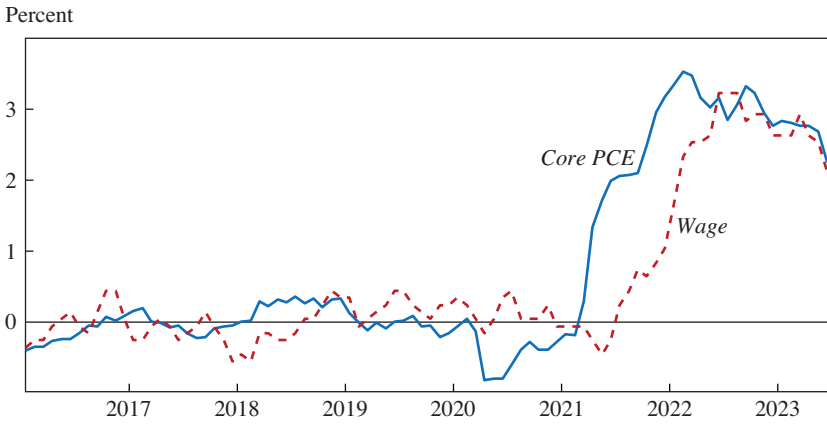


Source: CPI data from Bureau of Labor Statistics, retrieved from FRED; and wage inflation data from Wage Growth Tracker, Atlanta Fed.

specifically, and as figure 1 makes clear, wage inflation is, on average, higher than price inflation (equivalently, the real wage displays an upward trend). When using a simple plot to ascertain the impact of a shock on both variables, it is important to subtract their respective means. This is shown in figure 2, which displays the US price and wage inflation net of their (pre-COVID-19) means. The picture that emerges is significantly different, with more limited evidence of persistently higher wage inflation than price inflation (both relative to trend) at the end of the sample period. In other words, there is no evidence of a tendency for the real wage to revert back to its initial trend. That caveat appears even stronger when one uses core Personal Consumption Expenditures (PCE) data to construct the series for price inflation, as illustrated in figure 3.

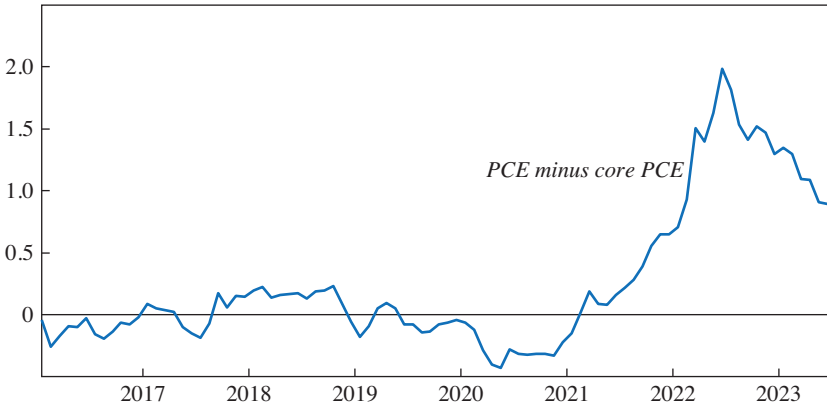
The resulting picture does not accord easily with the model simulations shown later in the paper, which imply trend reversion of the real wage. A possible explanation for the apparent absence of such trend reversion in the data is that the shock experienced by the US economy may have warranted a permanent fall in the real wage. Through the lens of the paper's model, this would be the case in the face of a permanent decline in the energy input endowment. Figure 4 displays some evidence consistent (if nothing else) with the hypothesis of a permanent supply shock: the log deviation between the PCE and core PCE indexes—which can be interpreted

Figure 3. Core PCE and Wage Inflation (Demeaned)



Source: Core PCE data from Bureau of Economic Analysis, retrieved from FRED; and wage inflation data from Wage Growth Tracker, Atlanta Fed.

Figure 4. Transitory or Permanent Shocks?



Source: Bureau of Economic Analysis, retrieved from FRED.

as a proxy for the relative price of noncore components (energy and food)—displays a seemingly permanent increase in the post-COVID-19 period relative to its stable pre-COVID-19 values.

A correct diagnosis of the forces behind the evidence is key to assess the challenges posed by wage developments in the near future, and in particular, by an eventual persistent above-trend wage inflation, possibly motivated by workers' resistance to seeing their real wage eroded. If the hypothesis of a permanent adverse supply shock is correct, that resistance should indeed be a source of concern since, ceteris paribus, it would be inconsistent with the attainment of the Federal Reserve's inflation target. Bringing back inflation to target would require, in that scenario, a recession strong enough to break the downward rigidity in real wages. The extension of the New Keynesian model allowing for real rigidities, developed in section V of the paper, would seem to provide the right framework for analyzing the options facing a central bank in that environment.

ON INFLATION AS CONFLICT As shown in the paper, aggregation of price-setting decisions in the continuous time version of a New Keynesian model yields the following expression for price inflation $\pi_t \equiv \dot{p}_t$:

$$\pi_t = \Lambda_p \int_t^{\infty} e^{-\rho(s-t)} \left[\left(w_s - p_s \right) - \left(mpl_s - \mu^p \right) \right] ds, \tag{1}$$

where w_s is the (log) average nominal wage, p_s is the (log) price level, mpl_s is the (log) marginal product of labor, and μ^p is the desired (or natural) price markup, assumed to be constant. Note that in contrast with equation (13) in the paper, I do not use demeaned variables, instead showing the constant term explicitly. Coefficient Λ_p, formally defined in the paper, is inversely related to the degree of price stickiness. Parameter $\rho > 0$ is the representative household's time discount rate.

Lorenzoni and Werning use equation (1) as a reference when putting forward their notion of inflation as conflict. Under that perspective, a rise in (price) inflation emerges when firms' real wage aspirations, defined by $mpl_s - \mu^p$, lie below actual real wages, either currently or anticipated. In that case, firms that get a chance to adjust their prices will tend to raise the latter, generating positive inflation.

A similar reasoning carries over to wage inflation, $\pi_t^w \equiv \dot{w}_t$, which is given by

$$\pi_t^w = \Lambda_w \int_t^{\infty} e^{-\rho(s-t)} \left[\left(mrs_s + \mu^w \right) - \left(w_s - p_s \right) \right] ds, \tag{2}$$

where coefficient Λ_w is inversely related to the degree of wage stickiness. Note that wage inflation is driven by current or anticipated gaps between workers' real wage aspiration, given by the (log) marginal rate of substitution augmented with the desired wage markup, $mrs_s + \mu^w$, and the actual average real wage $w_s - p_s$.

Accordingly, whenever firms' and workers' real wage aspirations are mutually inconsistent, this will necessarily be manifested in either price or wage inflation (or both), thus leading to the authors' view of inflation as conflict. In particular, whenever the path of real wages lies below that of workers' wage aspirations but above that of firms' corresponding aspirations, the implied upward pressure on wages and prices will reinforce each other, giving rise to a wage-price spiral, the focus of the paper.

The previous interpretation of inflation as conflict raises a number of questions, at least when applied to the New Keynesian model. In particular, I believe it gives a somewhat misleading impression about *individual* firms' motives. What drives the pricing decisions of an individual firm is the maximization of its value, which under the model's assumptions is attained by keeping its markup as close as possible (on average) to the optimal (flexible price) markup μ^p. In order to set its price optimally, the individual firm only needs to know its own nominal marginal cost, current and expected. Once that path is known, the real wage of its workers (defined relative to the entire consumption basket) is not of relevance to the price-setting firm. In particular, it does not care if the real wage of its workers goes up as a result of a reduction in other firms' prices.

The markup-based interpretation of an individual firm's motives, which can be read directly from the first-order condition associated with its optimal price-setting decision, is also reflected in inflation equation (1) once we rewrite it as follows:

$$\pi_t = \Lambda_p \int_t^\infty e^{-\rho(s-t)} \left[\mu^p - \left\{ p_s - \left(w_s - mpl_s \right) \right\} \right] ds$$

$$= \Lambda_p \int_t^\infty e^{-\rho(s-t)} \left(\mu^p - \mu_s^p \right) ds,$$

where $\mu_s^p \equiv p_s - (w_s - mpl_s)$ is the average price markup (with $w_s - mpl_s$ measuring the average marginal cost). Similarly, for wage inflation one can write

$$\pi_t^w = \Lambda_w \int_t^\infty e^{-\rho(s-t)} \left(\mu^w - \mu_s^w \right) ds,$$

where $\mu_s^w \equiv (w_s - p_s) - mrs_s$ is the average wage markup. Through this lens, price and wage inflation have a natural interpretation as the result of misalignments between actual and desired price and wage markups, respectively, and the consequent decisions by firms and workers in order to minimize those misalignments (at least in an expected sense), when allowed to do so.[1]

To be clear, the model is what it is, independent of the stories one can tell about its underlying mechanisms, and the wage-price block of the authors' model is fully standard. But to the extent that those stories help us understand the workings of the model, I can see two advantages of the interpretation based on markup misalignments relative to inflation as conflict advocated by the authors. First, while inflation is driven by deviations of a particular variable from some reference target in both cases, under the authors' interpretation that variable is the real wage whose target varies continuously over time and may even be nonstationary. By contrast, under the interpretation I prefer, the driving variable is the markup whose target is constant under standard assumptions. Second, as argued above, the markup misalignment interpretation seems to capture better the perspective of individual firms when making their price-setting decisions.

Finally, it is worth noting that the markup-based interpretation of inflation also provides a simple narrative for wage-price spirals. To see this, consider an adverse supply shock which raises firms' marginal costs and, as a result, lowers price markups relative to target. Firms that have a chance to adjust their prices will, on average, raise them, thus generating positive price inflation. Workers' real wages will be eroded as a result, thus lowering their average wage markup relative to target and inducing nominal wage increases among those workers who have a chance to reset their wage. The resulting wage inflation will in turn raise firms' marginal costs, leading to a second round of upward price adjustments, and so on.

CONFLICT INFLATION AND WAGE-PRICE SPIRALS Lorenzoni and Werning introduce the concept of conflict inflation as a component of price and wage inflation that results from a conflict between the wage aspirations of firms and workers. Formally, they define conflict inflation as follows:

$$\Pi_t^C \equiv \frac{\Lambda_p \Lambda_w}{\Lambda_p + \Lambda_w} \int_t^\infty e^{-\rho(s-t)} \left[\left(mrs_s + \mu^w \right) - \left(mpl_s - \mu^p \right) \right] ds, \tag{3}$$

1. See, for example, Galí (2015) for a textbook treatment of the New Keynesian model that stresses this interpretation.

where, once again, I am making explicit the constant terms in the expression. Note that conflict inflation is a discounted integral of current and future gaps between workers' real wage aspirations, $mrs_s + \mu^w$, and the corresponding aspirations for firms, $mpl_s - \mu^p$. A central theme in the paper is the connection between conflict inflation, thus defined, and the presence of a wage-price spiral. What is the nature of that connection?

Note that by combining equations (1) and (2) with the above definition of conflict inflation, one can show:

$$\Pi_t^C = \alpha\pi_t + \left(1-\alpha\right)\pi_t^w, \tag{4}$$

where $\alpha \equiv \dfrac{\Lambda_w}{\Lambda_w + \Lambda_p} \in \left[0,1\right]$. In words, conflict inflation can be expressed as a particular weighted average of price inflation and wage inflation, with the weight of each variable increasing in its relative stickiness.

A straightforward algebraic manipulation of equation (4) allows the authors to obtain the following expressions for price and wage inflation:

$$\pi_t = \Pi_t^C - \left(1-\alpha\right)\dot{\omega}_t \text{ and} \tag{5}$$

$$\pi_t^w = \Pi_t^C + \alpha\dot{\omega}_t, \tag{6}$$

where $\dot{\omega}_t \equiv \pi_t^w - \pi_t$ is the change in the real wage. Equations (5) and (6) motivate the authors' intended connection between conflict inflation and wage-price spirals, since Π_t^C can be interpreted, in their words, as the "underlying common component of price and wage inflation due to the gap between the aspirations on the two sides of the market."

However, establishing a rigorous connection between conflict inflation and wage-price spirals requires a formal definition of the latter. What is a wage-price spiral, after all? How can one measure its intensity?

While macroeconomists likely share at least a vague notion of what a wage-price spiral is, as far as I can tell there is no consensus on a formal definition of that phenomenon.[2] A possible definition, and one that the authors adhere to in several instances throughout their paper, is an episode

2. A recent paper by International Monetary Fund economists (Alvàrez and others 2022) seeks to identify wage-price spiral episodes throughout history. They use as a definition the observation of three successive quarters with accelerating price and wage inflation.

in which both price and wage inflation are positive.[3] Note, however, that conflict inflation would not seem to be a good indicator of the intensity of a wage-price spiral under such a definition, for any positive value of conflict inflation is consistent with wage and price inflation values of different sign.[4]

Furthermore, it is not obvious why any arbitrary weighted average of price and wage inflation (defined by a weight α different from $\frac{\Lambda_w}{\Lambda_w + \Lambda_p}$) could not also be thought of as a plausible wage-price spiral indicator, since equations (5) and (6) would also hold for that alternative measure. That measure, however, would bear no simple relation with conflict inflation.

So the question remains: what makes the particular weighted average of price and wage inflation defined by equation (4) with $\alpha = \frac{\Lambda_w}{\Lambda_w + \Lambda_p}$ (and which corresponds to conflict inflation) special or particularly desirable as a measure of wage-price spirals?

To address that question, the authors first propose a formal measure of the intensity of wage-price spirals, which they refer to as "spiral inflation." Formally, they define spiral inflation (in response to a shock at time zero) as:

$$\Pi_0^S = \int_0^\infty \pi_s ds,$$

that is, the cumulative change in price inflation. To the extent that the shock under consideration does not have a long-run effect on the real wage (as assumed by the authors), it follows $\int_0^\infty \pi_s ds = \int_0^\infty \pi_s^w ds$, that is, the cumulative change in wage inflation must equal that of price inflation, with their common value corresponding to spiral inflation, the authors' proposed wage-price spiral indicator.

Next the authors move on to show that, in the particular case that conflict inflation decays exponentially, spiral inflation will be proportional to conflict inflation. To see this, note that

3. More generally, one could define a wage-price spiral episode as one displaying price and wage inflation above their corresponding steady-state values. In the authors' model, those steady-state values are zero by assumption.

4. On the other hand, positive conflict inflation is a necessary condition for a wage-price spiral under that proposed definition. As the authors argue, however, positive conflict inflation necessarily implies positive cumulative price and wage inflation through the adjustment to the steady state, under certain assumptions.

$$\begin{aligned}\Pi_0^S &= \int_0^\infty \pi_s ds \\ &= \int_0^\infty \Pi_s^C ds - (1-\alpha)\int_0^\infty \dot{\omega}_s ds \\ &= \int_0^\infty \Pi_0^C e^{-\delta s} ds - 0 \\ &= \frac{1}{\delta}\Pi_0^C,\end{aligned}$$

where δ is the rate of decay of conflict inflation and $\int_0^\infty \dot{\omega}_s ds = 0$ follows from the stationarity of the real wage. The previous finding is interpreted by the authors as implying that "conflict inflation at date zero fully captures the underlying forces that lead to a protracted period of joint price and wage inflation," thus establishing the desired connection between conflict inflation and wage-price spirals.

The interest of the previous result notwithstanding, it is important to point out some caveats. First, the proportionality between spiral inflation Π_0^S and conflict inflation Π_0^C holds in the particular case of exponential decay, but it will not hold more generally. While such an exponential decay may be supported by an appropriate choice of monetary policy, it is generally not a property of the equilibrium. Furthermore, the coefficient of proportionality between the two variables depends on the rate of decay, which will not be invariant to the persistence of the shock or the policy rule in place. Accordingly, similar readings of conflict inflation at different points in time (or for different economies) may correspond to different levels of spiral inflation. Second, the tight relation between conflict inflation and spiral inflation hinges on the assumption of a stationary real wage, which is needed for $\int_0^\infty \dot{\omega}_s ds = 0$ to hold. Accordingly, the simple relation between spiral inflation and conflict inflation will vanish in the face of shocks with permanent effects on the real wage. Third, and perhaps most important, even in the case of a stationary real wage, the link between spiral inflation and conflict inflation uncovered above holds at time zero, that is, the time of the shock, when the real wage is still at its steady-state level, but it fails to do so on an arbitrary period $t > 0$ when that variable is away from the steady state, for in that case $\int_0^\infty \dot{\omega}_s ds \neq 0$.

CONFLICT, SPIRALS, AND THE DESIGN OF MONETARY POLICY Section IV of Lorenzoni and Werning's paper revisits the problem of optimal policy in the face of supply shocks. Given that the analysis of optimal policy in the New Keynesian model with staggered prices and wages, tracing back to

Erceg, Henderson, and Levin (2000), is generally well understood, some of the authors' findings are not entirely novel, though they are recast here in terms of conflict inflation and, more generally, they are related to the notion of a wage-price spiral. In particular, there are two well-established results in the literature on optimal policy in the model in Erceg, Henderson, and Levin (2000).[5] First, there exists a specific weighted average of wage inflation and price inflation, referred to as "composite inflation" in Galí (2015), for which the divine coincidence holds, that is, full stabilization of that variable implies full stabilization of the output gap. Second, there is a knife-edge parameter configuration for which the optimal policy calls for a full stabilization of the output gap and, hence, of composite inflation. More generally, and for a broad range of parameter values, such a policy is nearly optimal.

The connection between the previous results and some of the findings in the paper becomes clear once we recognize that the weighted average defining conflict inflation in equation (4) matches exactly the one that defines composite inflation in the existing literature. In particular, the symmetric case considered by the authors in their example 1 corresponds to the knife-edge case referred to above, while examples 2 and 3 can be viewed as an illustration of the near optimality of stabilization of the output gap more generally as reflected in the tiny response of that variable (once the scale of the plot is taken into account) under the optimal policy, as displayed in figures 10 and 11 in the paper.

Beyond the connection with the existing literature, the authors' analysis uncovers some results that shed light on a number of issues and that, in my opinion, are not sufficiently stressed in the paper.

First, the authors derive the second-order approximation to the welfare losses for the case of continuous time. The resulting expression is similar to the one for the discrete time case, originally derived in Erceg, Henderson, and Levin (2000). It is worth noting a difference, not emphasized by the authors, related to their use of a CES production function: the coefficient on the output gap Φ_y is inversely related to the elasticity of substitution between energy and labor. Thus, ceteris paribus, a low value for that elasticity will be associated with a higher weight on output gap stability in the central bank's loss function. That result, in a model in which the divine coincidence does not hold, is of great interest and its implications would seem to deserve some further discussion.

5. See proposition 3.9 and section 4.4 in Woodford (2004) and section 6.4.3 in Galí (2015).

Second, the authors note the following result, which follows from equation (4): with a zero output gap (and, hence, zero conflict/composite inflation), the adjustment in the real wage never requires positive inflation for *both* wages and prices. A slight generalization of that result, based on the near-optimality findings mentioned above, would run as follows: the fact that the optimal policy involves, at most, tiny deviations of conflict inflation from zero, rules out non-negligible positive inflation for both wages and prices as an optimal outcome. In their example 3, the authors uncover an instance of coexistence of positive wage and price inflation for a very brief period during the adjustment, but one should note that wage inflation is almost zero during that brief episode.

Under a definition of wage-price spirals as episodes with (non-negligible) positive inflation in wages and prices, the previous discussion would establish an interesting connection between optimal policy and the subject that is the focus of this paper, namely, the observation that wage-price spirals are (almost) always suboptimal. But, as discussed above, this is not the definition of wage-price spirals adopted by the authors, who instead focus on the concept of spiral inflation as an indicator of the intensity of wage-price spiral episodes. Unfortunately, the usefulness of spiral inflation in the context of the authors' optimal policy exercise is limited, since the real wage is permanently affected by the shock considered, implying that the mapping between conflict and spiral inflation is lost. In fact, under the optimal policy, and given the discussion above, we have

$$\Pi_0^S \simeq -\left(1-\alpha\right)\int_0^\infty \dot{\omega}_s ds,$$

which may take a large positive value in response to an adverse supply shock even if wage inflation and price inflation co-move negatively during the adjustment period (as in the three examples considered). It is clear that, in that instance, spiral inflation would not be a good indicator of a wage-price spiral.

ADAPTIVE EXPECTATIONS AND REAL RIGIDITIES Section V departs from the standard model in Erceg, Henderson, and Levin (2000) by exploring the implications of two potential sources of inertia, namely, a form of adaptive expectations that implies de-anchoring and the presence of real rigidities. The former is modeled by assuming that firms and workers expect constant inflation at all horizons (at a level that may be different from the steady state, thus the interpretation as a form of de-anchoring), with that variable adjusting slowly in response to variations in realized inflation. The latter

assumes that the real wage targets of workers and firms adjust sluggishly in response to changes in mrs_s and mpl_s.

Lorenzoni and Werning show that the introduction of de-anchoring leads to both greater inertia and higher persistence in both price and wage inflation relative to the baseline model, as a result of a strong underlying wage-price spiral mechanism. That prediction is enhanced when real rigidities are added.

Unfortunately, the authors do not carry out an analysis of optimal policy using the modified model. I believe it would be interesting to explore whether the two sources of inertia considered in this section could overturn the result derived for their baseline model, regarding the impossibility of non-negligible positive inflation coexisting for both wages and prices as an optimal outcome. I hope the authors (or someone else) undertake that analysis in future work.

Here is a minor quibble I have on this section: when considering the calibration with real rigidities (the second source of inertia), the authors maintain the assumption of adaptive expectations (the first source of inertia), but they lower the setting of γ from 1 to 0.1, which is justified on the grounds that "inflation expectations play a more limited role." This may be somewhat confusing to the reader since, as far as I understand, lowering γ makes inflation expectation even more sluggish (and thus further from rational expectations than in their first exercise where they only considered adaptive expectations as a source of inertia). In any event, I believe the authors should have gone back to rational expectations when studying real rigidities, in order to insulate the independent role played by this second source of inertia.

As a final comment, I would encourage the authors to discuss the connection between the two sources of inertia and wage indexation, a feature that is often incorporated in estimated versions of the standard model.[6] Wage indexation is typically modeled by having the nominal wages that are not re-optimized to be adjusted automatically in proportion to past price inflation. That mechanism is a source of real wage rigidity whose implications would be worth contrasting with the type of real rigidity assumed by the authors.

CONCLUDING REMARKS Recent price and wage developments in the United States and other advanced economies have rekindled fears of a wage-price spiral that may hinder central banks' efforts to control inflation. Lorenzoni and Werning's paper seeks to understand those developments through the lens of a New Keynesian model with sticky prices and wages. The first

6. See, for example, Smets and Wouters (2007).

challenge is to come up with an operational definition and measure of a wage-price spiral. The authors' proposed measure, spiral inflation, seems to be useful under certain conditions but not generally. The authors also explore the usefulness of conflict inflation, a concept they introduced in earlier work (Lorenzoni and Werning 2022), in accounting for wage-price spirals, and its connection with spiral inflation. In the context of the New Keynesian model, conflict inflation turns out to coincide with the particular weighted average of price and wage inflation (composite inflation), the stabilization of which implies the stabilization of the output gap; thus, conflict inflation inherits all the normative implications associated with composite inflation. Furthermore, conflict inflation is shown to be proportional to spiral inflation under certain conditions. In my discussion, I have raised some caveats about the usefulness of both conflict inflation and spiral inflation to help us understand and measure wage-price spirals. That skepticism notwithstanding, I found the paper to be thought-provoking and insightful along many dimensions. The likely inefficiency of wage-price spirals is an implication of their analysis that I found particularly interesting. It would be interesting to explore the type of changes in the environment that would allow that result to be overturned. An analysis of the normative implications of the sources of inertia introduced by Lorenzoni and Werning would seem to be a natural starting point in that endeavor.

REFERENCES FOR THE GALÍ COMMENT

Alvàrez, Jorge A., John C. Bluedorn, Niels-Jakob H. Hansen, Youyou Huang, Evgenia Pugacheva, and Alexandre Sollaci. 2022. "Wage-Price Spirals: What is the Historical Evidence?" Working Paper 2022/221. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/11/11/Wage-Price-Spirals-What-is-the-Historical-Evidence-525073.

Erceg, Christopher J., Dale W. Henderson, and Andrew T. Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46, no. 2: 281–313.

Galí, Jordi. 2015. *Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications*, 2nd edition. Princeton, N.J.: Princeton University Press.

Lorenzoni, Guido, and Iván Werning. 2022. "Inflation Is Conflict." Working Paper. https://economics.mit.edu/sites/default/files/inline-files/conflict%20inflation_0.pdf.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97, no. 3: 586–606.

Woodford, Michael. 2004. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, N.J.: Princeton University Press.

COMMENT BY

AYŞEGÜL ŞAHIN The onset of the COVID-19 pandemic in early 2020 led to a brief yet deep economic downturn. Following a significant decline in economic activity, the economy experienced a resurgence, accompanied by an abrupt and unexpected rise in inflation. After lying dormant for two decades, inflation surged, with the Consumer Price Index (CPI) climbing from 1.4 percent in January 2021 to 8.9 percent in June 2022. The evolution of wage inflation followed a different pattern. The Employment Cost Index increased at a lower pace than the price inflation initially and real wages declined. Most recently, as price inflation declined, wage growth surpassed price inflation, resulting in a boost in real wages. The rise in real wages triggered concerns about a potential wage-price spiral, which may impede the return of inflation to its target level of 2 percent.

Lorenzoni and Werning provide a careful examination of price and wage inflation dynamics through the lens of the New Keynesian framework. Their analysis yields several important insights about the drivers and consequences of the recent high inflation episode. This comment reviews and interprets Lorenzoni and Werning's findings and suggests extensions for future research.

FRAMEWORK The authors consider a New Keynesian framework with both wage and price rigidities. An important addition to the standard model is a nonlabor input (X) with a flexible price and inelastic supply. This input, X, is the second input to production besides labor L. It broadly captures supply chain disruptions and the rise in the price of energy and raw materials reflecting pandemic-related factors that adversely affected production. An important assumption is the low substitutability between X and L, which is important for the initial surge in inflation. This is because when demand increases, the price of the nonlabor input X rises, leading to scarcity. Consequently, the marginal product of labor (MPL) declines, given the low substitutability between X and L. This scarcity contributes to a rise in noncore inflation, creating a distributional tension between workers and firms, potentially initiating a wage-price spiral. Notably, real wages initially decrease as price inflation picks up. The key takeaway from these dynamics is that the fact that nominal wage growth is currently exceeding price inflation could be given an optimistic interpretation. In particular, it might be interpreted as a sign of real wages going back to trend and not necessarily as a concern of an ongoing wage-price spiral.

Key conditions that the framework requires to match the price and wage dynamics since 2021 are summarized in proposition 2 in the paper. When an economy satisfies the condition stated in proposition 2, the authors refer

to it as a "supply-constrained" economy. This condition is met if some key assumptions are satisfied, specifically: X is inelastically supplied with flexible price, which allows its price to adjust rapidly; X plays a significant role in production with a high share (denoted as s_X) and acts as a complement to labor, characterized by low substitutability (ϵ); and wages are relatively more rigid than prices ($\Lambda_w < \Lambda_p$).

To summarize, the elasticity of substitution between the nonlabor input and labor, along with the relative rigidity of wages and prices, plays a crucial role in the joint dynamics of wages and prices. If the nonlabor input is less important in production and can be readily substituted by labor, the increase in inflation would be subdued. Moreover, the rigidity of wages in comparison to prices significantly influences the joint dynamics of price and wage inflation.

These key parameters that are highlighted in proposition 2 are likely to vary across different sectors of the economy. More specifically, goods-producing and service-providing sectors use different production technologies. The literature finds complementarity between intermediates, which supports the low elasticity of substitution assumption for the goods sector.[1] However, the elasticity of substitution is likely to be higher in the services sector, and wage rigidities are less likely to play an important role for services since labor turnover has been very high in the recent period. That is why the framework in the paper is likely to be more relevant for accounting for inflation dynamics in the goods sector.

GOODS AND SERVICES INFLATION Examining the goods-producing and service-providing sectors would be useful for digging deeper into inflation dynamics since the initial surge in the US inflation was almost solely driven by goods inflation. The pickup in services inflation has also been significant, but it has been more modest, and it lagged inflation in the goods sector as shown in figure 1. This is a reversal of the typical inflation dynamics in the last twenty years, which were characterized by pro-cyclical services price inflation and essentially zero goods price inflation over the past ten years.

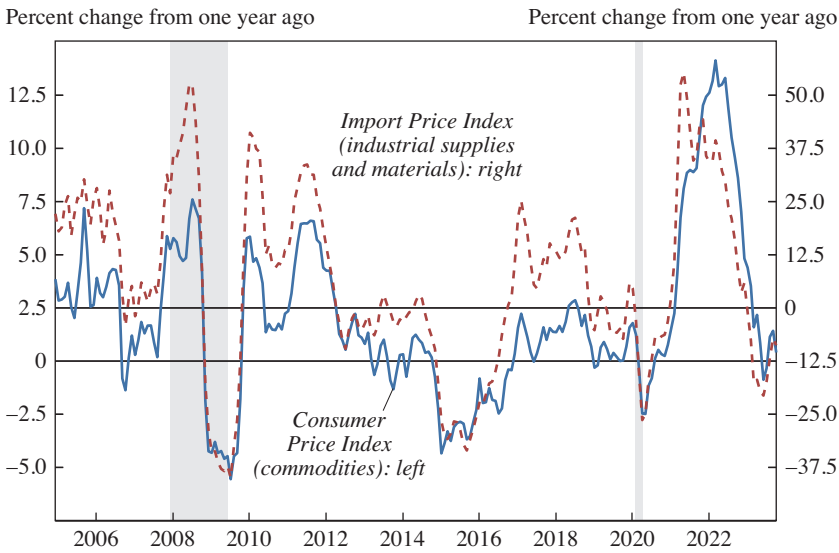
An important factor that is often cited for the surge in goods prices is supply chain bottlenecks. Figure 2 shows that the price of industrial supplies and materials has risen sharply, increasing by more than 50 percent at the onset of the pandemic. This increase coincided with the emergence of goods inflation and is often referred to as the main driver of a rise in prices.[2]

1. See, for example, Boehm, Flaaen, and Pandalai-Nayar (2019).
2. See, for example, Amiti and others (2023).

Figure 1. Consumer Price Indexes for Commodities and Services

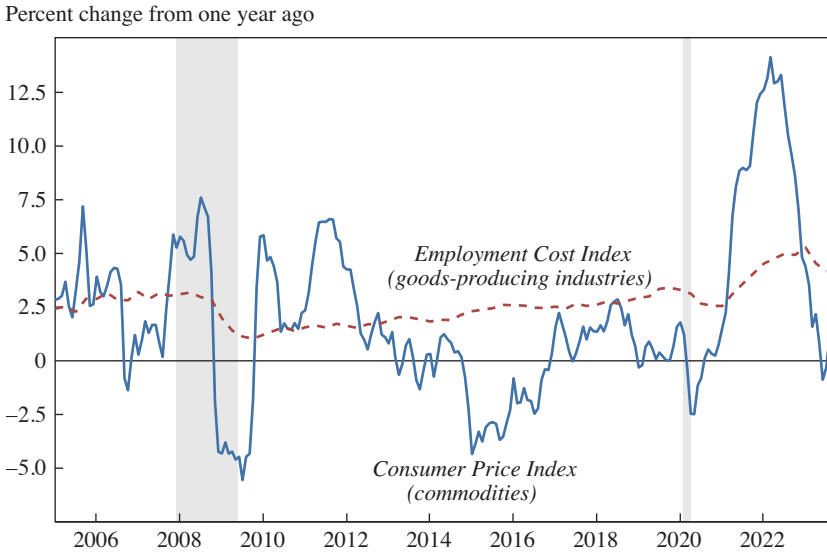
Source: Bureau of Labor Statistics, series CUSR0000SAC and CUSR0000SAS, retrieved from FRED.
Note: Consumer Price Indexes are for all urban consumers, US city average. Shaded areas indicate US recessions.

Figure 2. Consumer Price Index for Commodities and Import Price Index (End Use) for Industrial Supplies and Materials



Source: Bureau of Labor Statistics, series CUSR0000SAC and IR1, retrieved from FRED.
Note: Consumer Price Index is for all urban consumers, US city average. Shaded areas indicate US recessions.

Figure 3. Consumer Price Index for Commodities and Employment Cost Index for Wages and Salaries for Private Industry Workers in Goods-Producing Industries



Source: Bureau of Labor Statistics, series CUSR0000SAC and CIS202G0000000000I, retrieved from FRED.

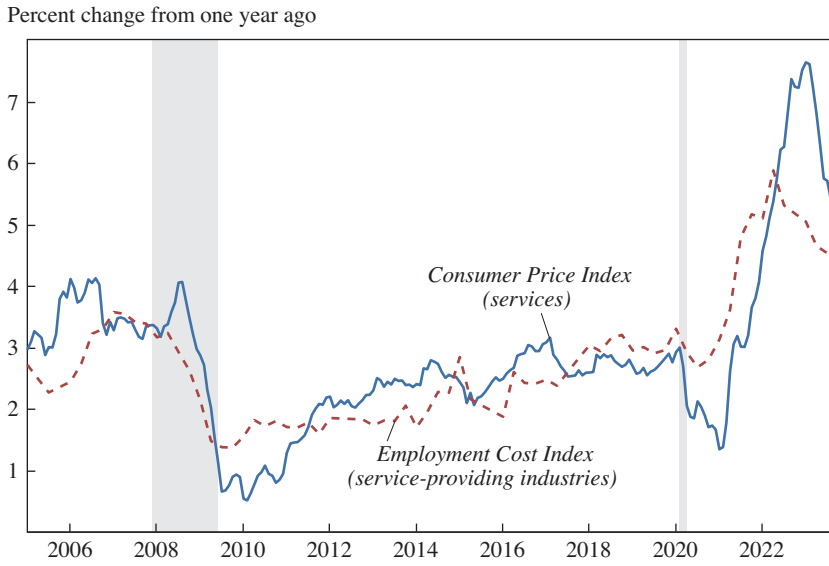
Note: Consumer Price Index is for all urban consumers, US city average. Shaded areas indicate US recessions.

These observations all support the authors' modeling choices. The introduction of the nonlabor input allows the model to account for the rise in prices of industrial supplies and raw materials. In addition, the assumption that it is inelastically supplied with a flexible price—which prevents their quantities from adjusting to relieve price pressures—mimics the supply chain disruptions related to the pandemic. Figure 3 shows the time series of price inflation in the goods sector along with wage growth in the sector. The evolution of price and wage inflation is similar to dynamics generated by the model.

However, price and wage inflation dynamics in the services sector look very different: wage inflation picks up before prices, and it also starts to retreat before prices, as shown in figure 4. While the model does a good job of accounting for joint price-wage dynamics in the goods sector, it is less applicable to the services sector.

WORKERS' AND FIRMS' ASPIRATIONS The authors define an interesting concept of inflation that they refer to as "conflict inflation." Fundamentally,

Figure 4. Consumer Price Index for Services and Employment Cost Index for Wages and Salaries for Private Industry Workers in Service-Providing Industries



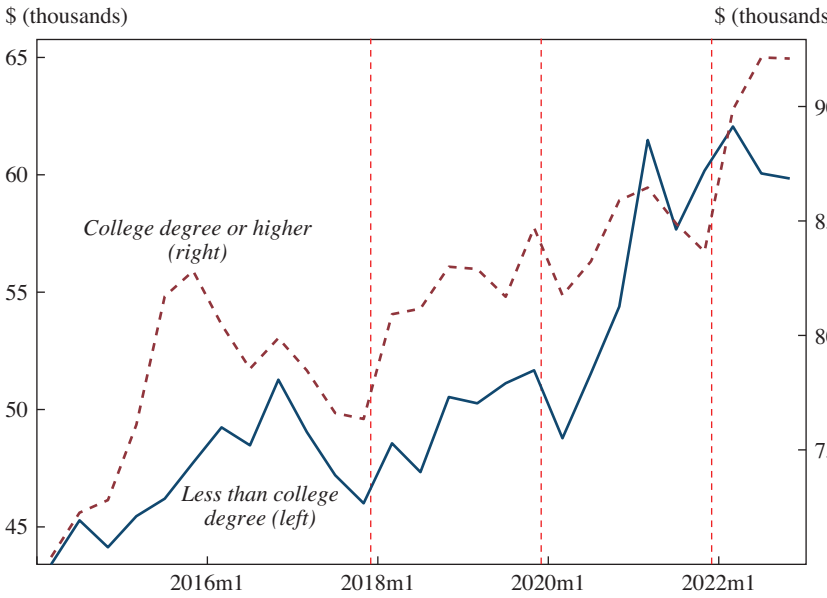
Source: Bureau of Labor Statistics, series CUSR0000SAS and CIS202S0000000000I, retrieved from FRED.

Note: Consumer Price Index is for all urban consumers, US city average. Shaded areas indicate US recessions.

the economic intuition behind the wage-price spiral mechanism lies in the divergence of views between workers and firms regarding the relative price of goods and labor, represented by the real wage W/P. When firms adjust nominal prices, they do so with a specific target for W/P in mind. However, workers may demand nominal wages with the aim of achieving a higher real wage. This conflict in aspirations leads to inflation in both prices and wages. This definition of a wage-price spiral emphasizes the disagreement or conflict as a key driver of inflation, as analyzed in a companion paper (Lorenzoni and Werning 2022).

While it is hard to measure the degree of disagreement between workers' and firms' aspirations, some new data sources provide us with some information regarding the evolution of these aspirations. A useful metric for summarizing workers' aspirations is the reservation wage of workers. The Federal Reserve Bank of New York's Survey of Consumer Expectations provides a measure of the reservation wages obtained from the following survey question: "Suppose someone offered you a job today in a line of

Figure 5. Reservation Wages by Educational Attainment



Source: Survey of Consumer Expectations, Federal Reserve Bank of New York.

work that you would consider. What is the lowest wage or salary you would accept (BEFORE taxes and other deductions) for this job?"

Figure 5 shows the reservation wages by educational attainment starting in 2014. Reservation wages started to rise for both workers with college education and those without in 2017, but the rise was much steeper for workers without a college degree. This increase is in line with the authors' characterization of the inflationary episode.

What about firms' aspirations or willingness to pay workers? Wages posted by employers with job openings provide a direct measure of firms' wage aspirations for the workers they plan on hiring. Crump and others (2022) utilize data from Burning Glass Technologies on posted job vacancies to examine posted wages. They find that, on average, posted wages for jobs with salaries below $75,000 grew at a rate of about 12 percent from 2019 to 2021 compared to about 8 percent from 2017 to 2019. The strong posted wage growth at lower salary positions over the last two years coincided with the stark rise in reservation wages of workers without college education.

Although there has been a rise in workers' wage aspirations, as indicated by their reservation wages, posted wages indicate that firms have met these

aspirations when posting job openings. Though these measures are only suggestive, they point to a less important role for conflict between firms and workers in driving inflation dynamics.

OTHER DEVELOPMENTS IN THE LABOR MARKET While the paper provides an intriguing explanation for wage and price inflation dynamics, several developments in the labor market point to the existence of other factors. Arguably, the most striking development in the labor market has been the so-called Great Resignation: the quits rate for employed workers reached 3 percent in 2021, almost 50 percent higher than in 2019.[3] Moreover, the Beveridge curve exhibited a wide loop and a vertical shift, unlike its commonly observed horizontal movements. The behavior of wages during the recovery from the pandemic recession also deviated from historical patterns. While high-wage workers typically experience faster wage growth during recoveries, leading to an increase in the wage gap between high- and low-wage workers, the opposite occurred after the pandemic, leading to wage compression, as documented by Autor, Dube, and McGrew (2023). One possibility is that the shift in worker preferences toward more flexible jobs coupled with a rapid recovery triggered an increase in quits by workers in search of more flexible job opportunities and put downward pressure on wages in high-amenity jobs, as argued by Bagga and others (2023). Under this interpretation, as reallocation from low- to high-amenity jobs subsides, job-to-job transitions could be more inflationary.

CONCLUDING REMARKS Lorenzoni and Werning provide a timely paper on an important topic with rich insights. They focus on conflict inflation as a key driver of the post-pandemic inflation surge. They also carefully study the interplay of supply chain disruptions and disagreement between workers and firms. For future research, exploring a multi-sector model and distinguishing between goods and services with different production technologies and degrees of wage and price rigidities, could provide valuable insights. Additionally, incorporating measures of workers' and firms' expectations about wages and prices would help improve the model's quantitative implications.

REFERENCES FOR THE ŞAHIN COMMENT

Amiti, Mary, Sebastian Heise, Fatih Karahan, and Ayşegül Şahin. 2023. "Inflation Strikes Back: The Role of Import Competition and the Labor Market." *NBER Macroeconomics Annual* 38.

3. Federal Reserve Bank of St. Louis, "Quits: Total Nonfarm," https://fred.stlouisfed.org/series/JTSQUR.

Autor, David, Arindrajit Dube, and Annie McGrew. 2023. "The Unexpected Compression: Competition at Work in the Low Wage Labor Market." Working Paper 31010. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31010.

Bagga, Sadhika, Lukas Mann, Ayşegül Şahin, and Giovanni L. Violante. 2023. "Job Amenity Shocks and Labor Reallocation." Working Paper. https://sadhikabagga.github.io/assets/pdf/Amenities/2023-11-23-Amenities.pdf.

Boehm, Christoph E., Aaron Flaaen, and Nitya Pandalai-Nayar. 2019. "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake." *Review of Economics and Statistics* 101, no. 1: 60–75.

Crump, Richard K., Stefano Eusepi, Marc Giannoni, and Ayşegül Şahin. 2022. "The Unemployment-Inflation Trade-off Revisited: The Phillips Curve in COVID Times." Working Paper 29785. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w29785.

Lorenzoni, Guido, and Iván Werning. 2022. "Inflation Is Conflict." Working Paper. https://economics.mit.edu/sites/default/files/inline-files/conflict%20inflation_0.pdf.

GENERAL DISCUSSION Jason Furman suggested that the authors look into empirical evidence on wage-price spirals outside of the motivating case of the United States to decide the usefulness of the proposed model. He pointed specifically to the recent, and quite different, experience in Europe: there was a larger shock to what the authors refer to as X, and real wages declined significantly, as did nominal wages after an initially modest response. This would have implications for the pro- and countercyclicality of wages in different situations. Referring to the way the authors define a wage spiral as following the logic of a conflict, Furman also proposed an approach where countries could be grouped according to institutions that maximize conflict and institutions that minimize conflict, looking at how their impulse responses differ.

Martin Baily contemplated what wage-setting model the authors had in mind, noting that there are different labor markets, including a very small part that is unionized and a part that is not. Given that many wages are set in a spot market or something close to it, how do aspirations fit into the picture? Baily was skeptical of the notion that workers with an aspiration for higher wages would simply be able to ask for them. To better gauge this part of the model, he advised the authors to look at how labor market institutions differ across countries or within the United States over time.

Guido Lorenzoni clarified that the central issue is not who sets the wages; rather, the general equilibrium problem is that once the nominal wage is set, the firm and the worker negotiating take the price of all other goods as a given. The firm can set the real wage in terms of the goods they produce

but not in terms of the goods that the worker consumes; thus, there is still a coordination problem to be solved.

Betsey Stevenson provided two examples that have arguably had an impact on workers' aspirations in terms of wages. Stevenson explained that many nurses quit their jobs during the labor shortage and signed up as travel nurses—often being assigned back to the same hospital but making a substantially higher wage. As a second example, she highlighted the fact that unionized workers have seen greater real wage declines than their non-unionized peers—an unusual development that is likely to leave unionized workers frustrated. Furthermore, Stevenson speculated that the widespread anger among workers, beyond those who have experienced falling real wages and despite a strong economy, will fuel expectations around the labor market for the next year or two.

David Romer asked the authors to elaborate on how their reinterpretation of inflation—conflict inflation—differs from standard accounts. In a standard New Keynesian model, for example, one could interpret an episode of inflation when output is above normal as a result of the set of monopolistically competitive firms having a form of conflict because they have mutually incompatible goals, as each would like its relative price to be above average. Romer also argued against focusing only on inflation expectations. Workers may demand higher wages not just because they expect inflation will be higher in the future but also because their real wages have fallen, pointing to the recent United Auto Workers strike as an example.[1] Romer therefore wondered whether aspirations might be a variable with a life of its own and, if added to the models, could provide additional insights.

Also raising options for expansions to the authors' model, Şebnem Kalemli-Özcan mentioned her own work using a production network in which one can identify both sectoral labor supply shocks—pointing specifically to the service sector as relevant here—as well as nonlabor, goods shocks.[2] To Furman's point, Kalemli-Özcan stated that the timing and

1. Reuters, "UAW to Expand Strike at Ford, General Motors," September 29, 2023, https://www.reuters.com/business/autos-transportation/uaw-expand-strike-ford-general-motors-2023-09-29/.

2. Cem Çakmakli, Selva Demiralp, Şebnem Kalemli-Özcan, Sevcan Yeşiltaş, and Muhammed A. Yildirim, "The Economic Case for Global Vaccinations: An Epidemiological Model with International Production Networks," working paper 28395 (Cambridge, Mass.: National Bureau of Economic Research, 2022), https://www.nber.org/papers/w28395; Julian di Giovanni, Şebnem Kalemli-Özcan, Alvaro Silva, and Muhammed A. Yildirim, "Pandemic-Era Inflation Drivers and Global Spillovers," working paper 31887 (Cambridge, Mass.: National Bureau of Economic Research, 2023), https://www.nber.org/papers/w31887.

intensity of the nonlabor goods, the labor supply, and labor demand shocks were very different across countries, noting as an example how there is growing consensus in the United Kingdom that the labor supply shock is of a permanent nature, that is, it is expected to permanently reduce the labor force.

Gerald Cohen proposed that thinking about inflation requires a high level of sophistication, including separating goods and services. To properly assess a model of inflation, we ought to investigate how the various markups from different parts of the economy are translated into the inflation numbers.

Iván Werning responded to the calls for a more sophisticated model by explaining that the intent was to generate a very stylized, general model. The authors wanted to capture the fact that the wages are sluggish in the simplest possible way. Werning further argued that the policy debate is often even simpler than the model they proposed, missing simple aspects of the inflation issue that the authors wanted to point to using their model. In terms of the shocks, Werning said that other types of shocks—a permanent one, for example—could easily be incorporated into the model, as could aspirations. He emphasized that they had not been omitted because the authors did not believe in them. Werning noted that the strength of their model is precisely the fact that it is very general, and the purpose of their paper is to provide a different perspective using a standard model, perhaps with a tweak to the parameters. Responding to questions about introducing aspirations into the model, Werning referred to a paper by Olivier Blanchard and Jordi Galí as a source of inspiration for authors' other work where they introduce the possibility that workers demand higher wages for reasons that go beyond nominal wage rigidities.

Lorenzoni explained that their model is completely compatible with a multi-sector approach in which, as we saw in the most recent episode, service inflation lags behind goods inflation, for example. In the paper, the simplest case with two sectors is presented: one produces goods and the other "produces" labor. But even in this simple case, the model provides the same, important intuition: different sectors react with a different lag in response to a shock, which gives rise to a sort of ripple effect that travels through the economy. Lorenzoni pondered the necessary preconditions for such a ripple effect to take place, noting that a price increase is not always enough to create a wave. But if we collectively were to lose faith in the stability of the unit of account, the ripple effect that follows would see the higher cost being passed along from the goods-producing sector to firms, and

then to consumers—the wage earners—who would negotiate for a higher wage. This is the source of conflict that the paper highlights.

Justin Wolfers made the point that the 6 percent private-sector unionization rate in the United States perhaps does not lend itself very well to the frame of the proposed model but would fit a labor market like that in Australia quite well.[3] Offering suggestions for future additions, Wolfers encouraged the authors to add a third player to their model: central banks. He was curious whether there would be distributional consequences as a result of central banks adopting an inflation-targeting regime as opposed to a nominal wage target.

Caroline Hoxby was struck by how the authors' findings seemed to have a clear analogue in the public finance literature. The shock in this case would be a tax reform, and the findings typically indicate a response that is quite fast for workers whose earnings depend on prices—a car dealer, for example. The response is significantly smaller for workers whose earnings depend on wages. This produces a strikingly similar pattern to figure 1 in the paper.

Michael Kiley offered a different perspective from what he interpreted as the authors' conclusion: wage-price feedback is currently limited, which suggests there is no cause for concern. Referring to the empirical literature on Phillips curves, Kiley highlighted two stylized facts. Wage-price feedback was limited from the 1990s to the 2010s but was much more pronounced in the 1970s and the 1980s. Kiley argued that the data tend to support that we are currently in a situation that more closely resembles the 1970s and the 1980s, citing his own work and suggesting that a lack of anchoring of inflation expectations or the really big shock we just experienced are the two most plausible explanations for the more apparent wage-price feedback we are seeing now.[4] While the data seem to support the latter explanation, Kiley emphasized that the real concern is the possibility that it is indeed inflation expectations that are drifting.

Benjamin Moll wondered if the authors could talk about how the results would change if the assumption in the model about perfect foresight was relaxed: for example, if workers were more myopic, and perhaps firms were more forward-looking than the workers.

3. US Department of Labor, Bureau of Labor Statistics, "Union Members—2022," news release, January 19, 2023, https://www.bls.gov/news.release/pdf/union2.pdf.

4. Michael T. Kiley, "The Role of Wages in Trend Inflation: Back to the 1980s?" Finance and Economics Discussion Series (Washington: Board of Governors of the Federal Reserve System, 2023), https://www.federalreserve.gov/econres/feds/the-role-of-wages-in-trend-inflation-back-to-the-1980s.htm.

Bringing us back to 2020, when real wages were higher and then started falling, Wendy Edelberg made the point that the effects of the supply shock—abstracting from demand—had to be absorbed somewhere: a drop in productivity and real income was inevitable. But how much of the cumulative real wage loss can be attributed to the supply shock?

BENJAMIN MOLL
London School of Economics

MORITZ SCHULARICK
Kiel Institute for the World Economy

GEORG ZACHMANN
Bruegel

The Power of Substitution: The Great German Gas Debate in Retrospect

ABSTRACT The Russian attack on Ukraine in February 2022 laid bare Germany's dependence on Russian energy imports and ignited a heated debate on the costs of a cutoff from Russian gas. While one side predicted economic collapse, the other side (ours) predicted "substantial but manageable" economic costs due to households and firms adapting to the shock. Using the empirical evidence now at hand, this paper studies the adjustment of the German economy after Russia weaponized gas exports by cutting Germany off from gas supplies in the summer of 2022. We document two key margins of adjustment. First, Germany was able to replace substantial amounts of Russian gas with imports from third countries, underscoring the insurance provided by openness to international trade. Second, the German economy reduced gas consumption by about 20 percent, driven mostly by industry (26 percent) and households (17 percent). The economic costs of demand reduction were manageable with the economy as a whole only experiencing a mild one-quarter contraction in the winter of 2022–2023 and then stagnating. Overall industrial production decoupled from production in energy-intensive sectors (which did see large drops) and declined only slightly. We draw a number of key lessons from this important case study about the insurance offered by access to global markets

Conflict of Interest Disclosure: The authors did not receive financial support from any firm or person for this paper, or from any firm or person with a financial or political interest in this paper. The authors are not currently an officer, director, or board member of any organization with a financial or political interest in this paper. The discussant, Tarek Hassan, is a cofounder of NL Analytics.

Brookings Papers on Economic Activity, Fall 2023: 395–455

and the power of substitution, specifically that supply shocks have dramatically smaller costs when elasticities of substitution are very low (but nonzero) compared to a truly zero elasticity.

> "Do we knowingly want to destroy our entire economy?"
>
> —BASF CEO Martin Brudermüller,
> *Frankfurter Allgemeine Zeitung*, March 31, 2022[1]

On March 7, 2022, less than two weeks after the Russian invasion of Ukraine, we published, jointly with a group of coauthors, a paper that addressed a seemingly simple question: what if the German economy was cut off from Russian gas? At that point, Germany imported about 55 percent of its gas consumption from Russia and relied on Russia for close to one-third of its total energy consumption (Bachmann and others 2022b). The "what if" question was intentionally framed in a way that allowed the cutoff to be the result of a German embargo or the result of an end to gas supplies initiated by Russia. The aim of the paper was to provide a compass for policymakers facing momentous decisions. How would the German economy cope with a sudden stop of energy imports from Russia? Would the likely result be a severe recession like during the global financial crisis or perhaps even a massive collapse in output and spiking unemployment comparable in its severity to the Great Depression of the 1930s? Or should we expect the economic costs to be more muted, that is, a more ordinary recession of the kind that the German economy had dealt with in the past and was well equipped to deal with in terms of the available policy space to cushion its impact?

Our answer at the time, based on key statistics about the German economy, relevant empirical estimates, and applied macroeconomic theory, was that an immediate emancipation from Russian energy was feasible and would entail substantial but manageable economic cost for the German economy. Our analysis foresaw an output cost in the first year following such a cutoff in the range of 1 to 3 percent relative to a no-cutoff baseline scenario, in line with previous recessionary episodes that the country had successfully dealt with. This prediction was highly controversial at the

1. The German company BASF is the largest chemical producer in the world and was heavily reliant on Russian gas until Russia cut off gas supplies to Germany in the summer of 2022. In the same interview, Brudermüller also warned that a cutoff from Russian gas "could bring the German economy into its worst crisis since the end of World War II and destroy our prosperity" (Brankovic and Theurer 2022).

time and triggered an intense public debate that culminated in the German chancellor warning of the "irresponsible" use of mathematical models for policymaking on a prime-time talk show.[2] Fearing catastrophic economic consequences of an end to Russian gas, the German government decided to keep importing rather than sanctioning it. Moreover, partly because of the fear of Russia retaliating by cutting off gas supplies, the German government was widely perceived to have taken a softer stance in offering support to the Ukrainian government and imposing other sanctions on Russia.

The Russian gas soon stopped flowing nevertheless. But it was Russia, not Germany or the European Union, that made the decision. Starting in June 2022, Russia drastically reduced gas supplies to Europe, in particular through the important Nord Stream 1 pipeline running directly from Russia to Germany in the Baltic Sea. Russia halted the Nord Stream 1 flows completely at the end of August 2022, and the pipeline was destroyed by underwater explosions four weeks later, resulting in a complete severance of Russian supplies to Germany.[3] One and a half years after the initial debate and a year after the final cutoff, this paper takes stock of what we have learned since then. We briefly review the original argument and the controversy it caused, but mainly focus on how the German economy coped with the actual severance of Russian gas supplies.

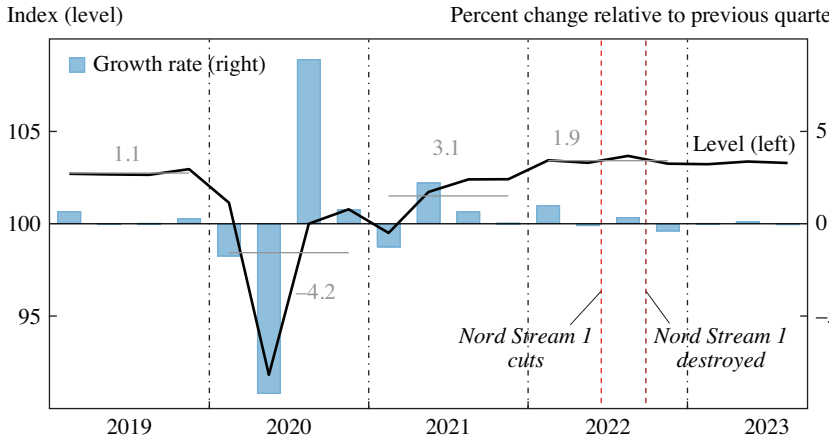
Prima facie, the evidence seems to support the original argument of the "what if" paper (Bachmann and others 2022b). Germany was partially cut off from Russian gas in June 2022 and completely in August 2022 but did not go into a deep depression. As shown in figure 1, Germany's gross domestic product (GDP) expanded by close to 2 percent for the entire year 2022 despite a circa 20 percent drop in gas consumption. In the fourth quarter of 2022, during the peak of the winter's heating season, the GDP contracted by 0.4 percent and stagnated thereafter, with growth in each of the first three quarters of 2023 close to 0 percent.[4] This outcome must be

2. *Anne Will* show with Chancellor Olaf Scholz on March 27, 2022; see https://benjaminmoll.com/Scholz/ for a transcript of excerpts with an English translation of Chancellor Scholz's comments. Key excerpt: "But they get it wrong! And it's honestly irresponsible to calculate around with some mathematical models that then don't really work."

3. BBC News, "Nord Stream 1: How Russia Is Cutting Gas Supplies to Europe," September 29, 2022, https://www.bbc.com/news/world-europe-60131520.

4. Of course, the observed evolution of German GDP is not directly comparable to a counterfactual prediction like ours that was relative to a no-cutoff baseline scenario holding other factors constant. The numbers for observed GDP have also been subject to repeated revisions. The data as of October 30, 2023 indicate that Germany experienced a technical recession (defined as two consecutive quarters of negative GDP growth) in the winter of 2022–2023 by the narrowest of margins, with GDP contracting by 0.4 percent and then 0.03 percent in the fourth quarter of 2022 and the first quarter of 2023.

Figure 1. Real GDP in Germany



Source: Destatis.

Note: The GDP data, seasonally and calendar adjusted, are from table 81000-0002 of the German National Accounts, available through Destatis at https://www-genesis.destatis.de/. The GDP level (left y-axis) is normalized to 100 in 2020:Q3, the quarter after the 2020 pandemic recession. Russia cut gas deliveries through the Nord Stream 1 pipeline substantially starting in mid-June 2022 (first to 40 percent, then 20 percent, "Nord Stream 1 cuts") and halted flows completely on August 31, 2022. The pipeline was destroyed on September 26, 2022 ("Nord Stream 1 destroyed").

compared to the estimates in studies financed by trade unions and business associations that foresaw output losses between 6 percent and 12 percent, with the most apocalyptic estimates due to Krebs (2022) and Prognos (2022), both of which predicted an output collapse of 12 percent, as well as Michael Hüther, who warned of "two and a half or three million additional unemployed" (IW 2022).[5] Overall, while the German economy is stagnating and faces substantial long-run headwinds, the direct economic costs

5. See Behringer and others (2022), Krebs (2022), and Prognos (2022). Even though counterfactual GDP predictions and the GDP time series are not directly comparable, it is clear that these dramatic counterfactual estimates between 6 percent and 12 percent have not come true. For example, given that GDP growth was close to zero over the 2022–2023 period, in order to believe a 12 percent GDP drop relative to a no-cutoff baseline scenario, one would have to believe that GDP would have grown at around 12 percent in the absence of a gas import stop, which is clearly absurd. For context, the Institut für Makroökonomie und Konjunkturforschung (IMK), which produced the report by Behringer and others (2022) is a union-financed think tank; the Krebs (2022) study was paid for by the German trade union federation, Deutscher Gewerkschaftsbund (DGB); and the Prognos study was paid for by a business association. See Bachmann and others (2022a) and Moll (2022) for a

of the end of Russian energy imports proved moderate and manageable, in line with the results of the original "what if" study.

In this paper, we have four main ambitions. First, we lay out the basic theoretical considerations regarding the economy's ability to adapt. One important and nonobvious point is that even very low elasticities of substitution are a powerful force for reducing the impact of a large input supply shock like the gas cutoff. While a Leontief production structure (i.e., the case in which elasticities are truly zero) implies drastic economic costs, specifically that production falls one-for-one with gas, even moderate substitutability mutes these costs considerably. The simplest illustration of this result uses a calibrated aggregate production function with an elasticity of substitution between gas and other inputs: in the Leontief case $\sigma = 0$, a 20 percent drop in gas supplies implies a 20 percent drop in production; however, when $\sigma = 0.05$, the corresponding output losses are only 2.7 percent, that is, going from $\sigma = 0$ to $\sigma = 0.05$ reduces the output loss by a factor of almost ten. The underlying logic is considerably more general, however, and extends to richer multi-sector models of supply chains like the model in Baqaee and Farhi (2024) used by Bachmann and others (2022b) to explore the importance of cascading effects in production (see section II). Intuitively, because the share of gas in production is small, even a small amount of substitutability is sufficient to overcome the gas input's bottleneck property. In the more complicated models, additionally, international trade plays an important role, specifically substitution of gas-intensive products via imports.

Second, we show how the German economy adapted to the end of Russian gas supplies. We track the consumption response of households and industries on the demand side and discuss the additional supply that replaced Russian gas. On the supply side, Germany was able to replace substantial amounts of Russian gas with imports from third countries, often taking advantage of the integrated European gas market, for example by importing US liquified natural gas (LNG) via LNG terminals in the Netherlands. On the demand side, the German economy reduced overall gas consumption by about 20 percent in the period July 2022 to March 2023

summary of studies conducted by other entities. For comparison, the German labor force was around 44 million people in 2022, so 2.5–3 million additional unemployed would have corresponded to an increase in the unemployment rate of more than 5 percent (data from World Bank, "Labor Force, Total – Germany," https://data.worldbank.org/indicator/SL.TLF.TOTL.IN?locations=DE). Michael Hüther is the head of industry-financed think tank Institut der Deutschen Wirtschaft (IW) Köln.

relative to previous years.[6] The largest contribution came from industry, which reduced its gas consumption by a striking 26 percent, whereas household gas consumption fell by a smaller but still impressive 17 percent. The online appendix complements these statistics by describing thirty-six concrete cases of substitution and adaptation by German firms and households.

We pay particular attention to the adjustment of the industrial sector to the gas cutoff. Much of the German debate in February and March 2022 centered around "cascading effects" in production, the idea that a cutoff from Russian gas would not only affect energy-intensive upstream sectors but also subsequently take down and "destroy" the entire industrial sector and economy with it—the quote by the BASF chemicals executive at the beginning of our paper is a good example of this line of argument. We therefore ask what sectors were most affected by the gas cutoff, and whether and to what extent it resulted in such cascading effects. While production in energy-intensive sectors like chemicals and glass production did see substantial cuts of up to 20 percent, we find no evidence of substantive cascading effects. To the contrary, we find that overall industrial production displayed a substantial decoupling from production in these energy-intensive sectors and was hardly affected. In an open economy with substitution possibilities, sharp declines in output in some upstream sectors do not necessarily lead to large contractions in downstream industries. At each point in the production network, substitution possibilities exist.

Third, we ask if Germany could have also withstood an earlier cutoff from Russian gas, as early as the end of March 2022, as advocated by some and hotly contested by others. A prominent line of thinking among the skeptics is that the additional five months from April to August, during which Germany continued to import and stockpile Russian gas, was decisive as it allowed the country to purchase enough Russian gas to increase storage capacity sufficiently to get through the following winter. By contrast, an immediate severance from Russian energy at the end of March 2022 would have resulted in storages running out in the middle of the winter as well as shortages and rationing, and an ensuing economic catastrophe.

We revisit this argument and show that Germany exited the 2022–2023 heating period with gas reserves that exceeded imports from Russia from April to August 2022. In other words, even in the scenario of a Russian supply cutoff at the end of March 2022, Germany would have had enough

6. The 20 percent overall demand reduction that we document is somewhat below other estimates in the literature. For example, Ruhnau and others (2023) find that gas consumption during the second half of 2022 was 23 percent below the temperature-adjusted baseline.

gas to make it through the following winter (assuming identical consumption). While actual observed gas storage levels were around 65 percent at the end of the 2022–2023 heating period, they would have still been around 25 percent even in the counterfactual scenario of an immediate cutoff. Moreover, as the March cutoff would have coincided with the end of the 2021–2022 heating period, the combination of gas imports from other countries and preexisting storage would have been sufficient to satisfy both industrial and household gas demand at any point in time. There would never have been a gas shortage at any point throughout the year, and German gas storage levels would have instead always exceeded a safety margin of around 25 percent. In other words, on the basis of this simple calculation, Germany would have been able to cope with an earlier embargo on Russian gas imports. The country's leaders likely overestimated the geoeconomic dependency on Russia and arguably opted for a more cautious policy toward Russia than was necessary.

Last, we briefly discuss the political economy of policy consulting and the role domestic lobbies have played in the process. We also look back critically and argue that Germany could have done more to help Ukraine at an earlier stage, and that there are important lessons for related cases in the future, such as China and Taiwan. Market economies have a tremendous ability to adapt, which we should not underestimate again.

The structure of this paper is as follows. We start with a short exposition of Germany's dependence on Russian gas before the Russian invasion of Ukraine and the events leading up to the eventual cutoff. Section II recaps the argument of the "what if" paper, specifically that substitution would be a powerful force toward lowering the costs of a gas cutoff. Section III discusses the adjustment that has taken place over the past year and benchmarks the development to the prediction of the model. Section IV asks whether an immediate disruption in April 2022 would have had much more severe consequences. Section V considers the role of "luck," specifically whether the 2022–2023 winter was particularly mild, as well as various other factors in global energy markets. Section VI discusses the main lessons from the debate for policy consulting and similar future episodes. Section VII concludes.

I. Background: Germany's Dependence on Russian Gas and the 2022 Gas Cutoff

Long ignored by German politicians, Germany's dependence on gas imports from Russia was exposed dramatically after the Russian aggression. How Germany became so dependent on Russian gas even though the Russian

government had weaponized its gas exports in the past (in particular against Eastern European countries like Ukraine), is a fascinating question for political scientists. A recent book by Bingener and Wehner (2023) provides an excellent analysis of the mix of political economy problems, industrial lobbying, naïveté, and outright corruption that led to this dependence. After Russia's attack on Ukraine, the question of economic dependence became one of acute geoeconomic relevance: to what extent were Germany's options to support Ukraine and take a tough stance on Russia compromised by the country's dependence on Russian gas?

Yet the European gas crisis started well before the Russian attack on Ukraine. Already in the summer of 2021, gas storages in Europe were not being refilled at the usual pace. Specifically, Russia's gas monopolist Gazprom controlled a number of storage facilities at the time, including Germany's largest one (Rehden), and purposely kept them almost empty. Russia gradually reduced gas supplies, withholding almost 20 percent of the usual pipeline flows it delivered to Europe in previous years. This led to sharply increasing gas prices from below €20 per MWh at the beginning of 2021 to a first peak of close to €100 per MWh in October, and a second peak of close to €150 per MWh in December 2021.[7] This gradual withholding of volumes by Russia went largely unnoticed by the media and did not enter into the public debate, likely in part due to the difficult access to gas flow data. Some commentators and so-called experts circulated various theories on technical, commercial, and legal reasons for the reduced flows, thereby preventing a sense of urgency among the policymakers and the public.

The start of the war had little direct impact on prices and volumes. However, when it became clear that Kyiv would not be taken in a few weeks and a coalition of Western countries formed that supported Ukraine and put substantial sanctions on Russia, Russia soon started further weaponizing its gas exports. To begin, the Russian president Vladimir Putin decreed on March 31, 2022,[8] that Gazprom would only receive payments for gas in Russian rubles. Even though this contradicted agreed contract terms and risked undermining financial sanctions, European policymakers were reluctant to offer clear guidance to their companies on this issue, likely due to the perceived importance of Russian gas imports for the functioning of Europe's economy. Subsequently, Gazprom stopped gas deliveries to Poland and Bulgaria for refusing to pay in rubles. Moreover, flows

7. Investing.com, "Dutch TTF Natural Gas Futures Interactive Chart," https://www.investing.com/commodities/dutch-ttf-gas-c1-futures-advanced-chart.

8. Reuters, "Putin's Decree on Russian Gas Purchases in Roubles," March 31, 2022, https://www.reuters.com/article/idUSL5N2VY5U7/.

through the Yamal pipeline (that passes Poland toward Germany) were also stopped by Russia based on claims of Polish sanctions against the pipeline company. In June 2022, Russia unilaterally limited gas flows through the Nord Stream 1 pipeline to 40 percent, then reduced them further to around 20 percent and eventually halted flows completely on August 31, 2022.[9]

These politically tense months between February and September 2022 were characterized by a Russian strategy to divide European unity, for example by selectively cutting gas supplies to specific countries while at the same time offering to Germany to open the newly built Nord Stream 2 pipeline so as to avoid the much-feared gas crisis.

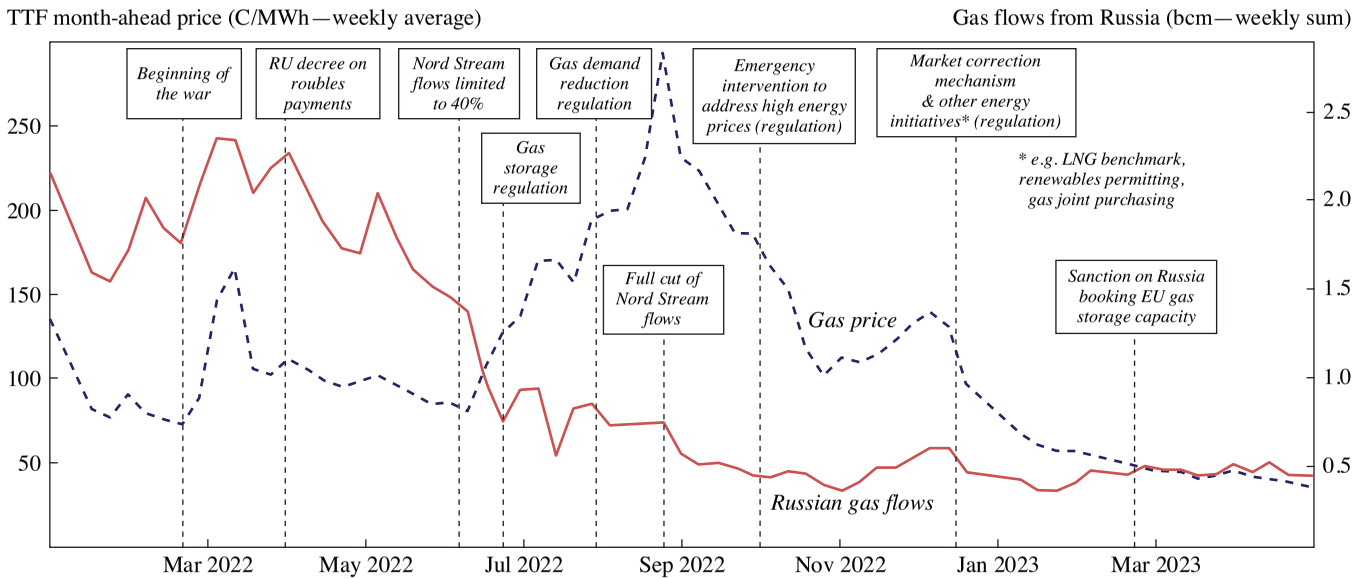
Finally, on September 26, 2022,[10] the two branches of Nord Stream 1 and one of the two branches of Nord Stream 2 were destroyed by underwater explosions in the Baltic Sea (with the actors unknown at the time of writing). The destruction of the Nord Stream pipelines ended this phase of uncertainty by substantially cutting Russian gas flows to Europe (routes via Turkey and Ukraine remained operational), in particular ending direct pipeline flows from Russia to Germany for good. While Germany imported more than half of its gas from Russia in 2021 (table 1), and this was expected to further increase with the planned opening of Nord Stream 2 at the beginning of 2022, the share of Russian gas fell to 0 percent by September 2022 (online appendix figure C.1). Figure 2 is reproduced from Gil Tertre (2023) and shows the key events over time.

The starting point of our "what if" paper was a summary of Germany's dependence on Russian energy at the beginning of the war in Ukraine (table 1). One energy input stood out: natural gas. In particular, data from 2021 showed that Germany imported more than half (55 percent) of its gas from Russia. Furthermore, Germany was much more dependent on natural gas than many other countries, with natural gas accounting for nearly a third of the overall energy mix.

9. Nina Chestney, "Russian Gas Flows to Europe Fall, Hindering Bid to Refill Stores," Reuters, June 16, 2022, https://www.reuters.com/markets/europe/russian-gas-flows-europe-fall-further-amid-diplomatic-tussle-2022-06-16/; Reuters, "Russia's Gazprom Tightens Squeeze on Gas Flow to Europe," July 26, 2022, https://www.reuters.com/business/energy/kremlin-nord-stream-1-turbine-be-installed-volumes-will-adjust-2022-07-25/; and Reuters, "Gazprom to Shut Down Nord Stream 1 Pipeline for 72 Hours," August 30, 2022, https://www.reuters.com/business/energy/nord-stream-1-nominations-fall-zero-aug-31-0200-cet-2022-08-30/.

10. Niha Masih, "Who Blew up the Nord Stream Pipelines? What We Know One Year Later," *Washington Post,* September 25, 2023, https://www.washingtonpost.com/world/2023/09/25/nord-stream-pipeline-explosion-update-russia-ukraine/.

Figure 2. Russian Weaponization of Gas Supplies and Gas Prices



Source: Reproduced from Gil Tertre (2023), copyright European Union.

Table 1. German Primary Energy Usage 2021

	Oil	*Natural gas*	*Coal*	*Nuclear*	*Renewables*	*Others*	*Total*
TWh	1,077	905	606	209	545	45	3,387
Percent	31.8	26.7	17.9	6.2	16.1	1.3	100
Percent (Russia)	34	55[a]	26	0	0	0	30

Source: Reproduced from Bachmann and others (2022b) with permission, copyright ECONtribute.
a. In 2020; already lower in 2021 and 2022.

In contrast to the other energy imports from Russia (oil and coal), it was also clear that Russian gas would be considerably harder to substitute with imports from third countries (like Norway or the Netherlands). This is due to German gas imports having been pipeline-bound, in particular from Russia via the Nord Stream and Yamal pipelines, and Germany at the time not having even a single terminal for importing LNG. The combination of Germany's large dependence on Russian gas and the difficulty in substituting this Russian gas with imports from other countries meant that we focused our analysis on the economic costs of a cutoff from Russian gas.

II. The Core Argument: The Power of Substitution

The core theoretical argument of the "what if" paper was that German firms and households would adapt to a cutoff of Russian gas supplies in ways that would ultimately reduce the economic impact. Producers would switch to other fuels or fuel suppliers and import products with high energy content, while households would cut their gas demand by turning down their thermostats. Importantly, elasticities of substitution that are very low, but nonzero, translate into much smaller economic losses than in the case of literally zero substitutability (i.e., Leontief production). Substitution along the supply chain and across producers would mean that macro elasticities are larger than micro elasticities. Cascading effects along the supply chain would be muted as opposed to "destroying" the economy's entire industrial sector.

Using the approaches we outline below, we argued that even in the case of a cold turkey import stop of Russian gas in March or April 2022, the economic costs would be substantial but manageable. Our analysis foresaw GDP and gross national expenditure (GNE) losses in the first year after such a cutoff in the range of 1–3 percent relative to a no-cutoff baseline scenario.

II.A. An Aggregate Production Function

To illustrate the power of substitution in a transparent fashion, we start by considering an extremely simple and purposely stylized setup. We assume that Germany produces output Y using natural gas G (which it imports from Russia) as well as other inputs X (like labor and capital), according to a constant elasticity of substitution (CES) aggregate production function

$$Y = \left(\alpha^{\frac{1}{\sigma}} G^{\frac{\sigma-1}{\sigma}} + (1-\alpha)^{\frac{1}{\sigma}} X^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \tag{1}$$

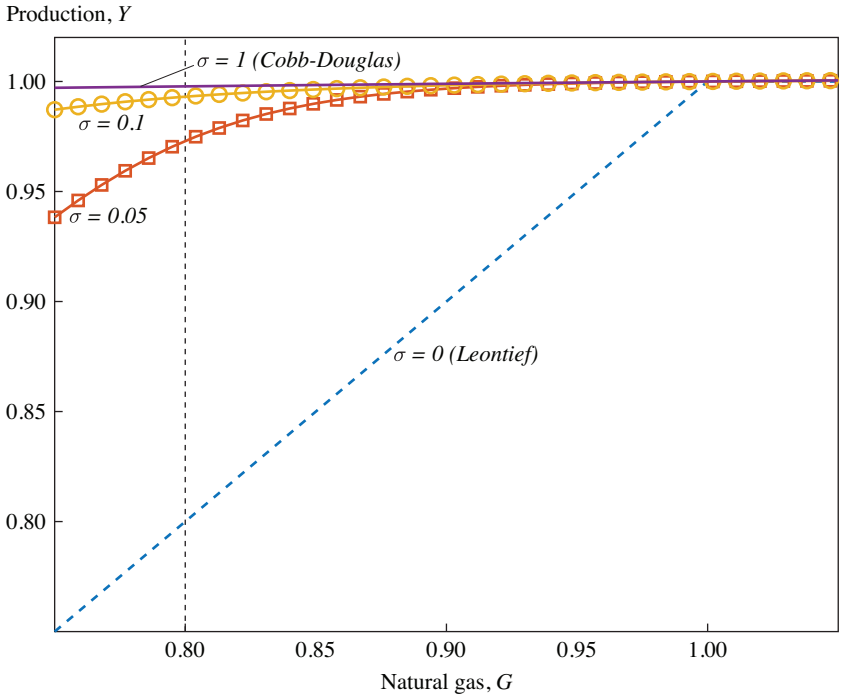
where $\alpha > 0$ parameterizes the importance of gas in production and $\sigma \in [0, \infty)$ is the elasticity of substitution between gas and other inputs. The goal is to assess the effect of a drop in gas supply G on production Y and how this depends on the features of the aggregate production function. The setup is, of course, extremely simplistic in that it only features two factors of production, no input-output linkages, and so on. However, as we discuss below, such an analysis can be a good approximation even in a much richer environment like the multi-sector model of Baqaee and Farhi (2024) used further below.

The following special cases show that, depending on the value of σ, the macroeconomic effects of a drop in gas supplies G are extremely different. The examples are complemented by figure 3, which plots production Y as a function of natural gas G for different values of the elasticity σ for a calibration described in Bachmann and others (2022b) in which the share parameter α equals 1 percent.[11]

A particularly useful special case is that of Leontief production, that is, exactly zero substitutability $\sigma = 0$, in which case equation (1) becomes $Y = \min\{G/\alpha, X/(1-\alpha)\}$. Starting from an initial optimum, a reduction in G implies that $Y = G/\alpha$ and hence $\Delta \log Y = \Delta \log G$. Therefore, if the elasticity of substitution is exactly zero, production Y drops one for one with gas supply G. This is illustrated by the dashed line in figure 3, which plots production Y as a function of G for the Leontief case. For example, a drop in gas supply of $\Delta \log G = -20\%$ implies a drop in production of $\Delta \log Y = -20\%$. Intuitively, the Leontief assumption means that, despite its small input

11. Bachmann and others (2022b) document that the share of natural gas consumption in German GNE is roughly 1 percent. This is also the share of gas imports in GNE because there is hardly any domestic production of natural gas.

Figure 3. Output Losses Following a Fall in Gas Supply for Different Elasticities of Substitution



Source: Authors' calculations.

share, gas is an extreme bottleneck in production: when energy supply falls by 20 percent, the same fraction (that is, 20 percent) of the other factors of production X lose all their value (their marginal product drops to zero), and hence production Y falls by 20 percent. Note that this output loss is completely independent of the input share α: with Leontief production, even a tiny input becomes an extreme bottleneck and takes down the economy one for one. That zero substitutability predicts production falling one for one with gas is much more general and is also true in multi-sector models with complex supply chains.

On the other extreme, the special case of Cobb-Douglas production with an unrealistically high elasticity of substitution of $\sigma = 1$ implies very small output losses. When $Y = G^{\alpha} X^{1-\alpha}$ we have $\Delta \log Y = \alpha \times \Delta \log G$ so that a 20 percent gas drop implies an output loss of only 0.2 percent ($1\% \times (-20\%) = -0.2\%$).

The most important conclusion, however, concerns intermediate cases with low but nonzero substitutability like $\sigma = 0.05$. The solid line with square markers in figure 3 plots the output losses for this case. It shows that the case with moderate but nonzero substitutability $\sigma = 0.05$ is very different from the Leontief case with literally zero substitutability $\sigma = 0$. For example, a 20 percent gas supply drop leads to an output loss of 2.7 percent rather than 20 percent, that is, going from $\sigma = 0$ to $\sigma = 0.05$ reduces the output loss by almost a factor of ten (at the same time, there is still substantial amplification relative to the 0.2 percent output loss in the Cobb-Douglas case $\sigma = 1$, again by roughly a factor of ten). Intuitively, because the input share $\alpha = 1\%$ is small, even a small amount of substitutability is sufficient to overcome the gas input's bottleneck property. In summary, while a Leontief production function predicts that production falls one for one with gas, even moderate substitutability implies much smaller losses.

For completeness and with an eye to other applications, we note that the value of the share parameter can also make a big difference. For example, suppose that $\alpha = 2\%$ rather than 1%. Then, in the Leontief case $\sigma = 0$, the output loss from a 20 percent gas supply drop is still 20 percent, that is, it is unaffected by the share parameter α. However, when $\sigma = 0.05$, $\alpha = 2\%$ implies an output loss of 4.5 percent rather than 2.7 percent. This point is particularly relevant in the context of other scenarios, for example, oil shocks (see section III.F) or China-Taiwan tensions.

Finally, it is worth noting that Bachmann and others (2022b) evaluated the effects of a gas cutoff not just on GDP but also on Gross National Expenditure (GNE). GNE, also known as "domestic absorption," is the economy's total expenditure defined as the sum of household expenditure, government expenditure, and investment, that is, $GNE = C + I + G$ in the GDP accounting identity $GDP = C + I + G + X - M$. GNE (rather than GDP) is the welfare-relevant quantity in many macroeconomic and trade models, including the Baqaee-Farhi model. One reason for focusing on GNE rather than GDP is that GDP may not pick up the terms of trade effect through which German consumers become poorer when the price of natural gas (an imported good) rises (Obstfeld and Rogoff 1995; Mendoza 1995).[12] Sinn (2022) misguidedly criticized the analysis of Bachmann and

12. Theoretically the effect is easiest to see in a small open endowment economy with an exogenously given relative price of exports to imports p (which is the country's terms of trade). Real GDP is given by the endowment and therefore not affected by fluctuations in the terms of trade p. However, consumption and welfare decline when the terms of trade p decline, an effect not picked up by real GDP.

others (2022b) for missing this effect even though GNE is not subject to this criticism.[13]

II.B. Macro Elasticities Are Larger than Micro Elasticities

The question under consideration in the great gas debate was the potential impact of a cutoff from Russian gas on the German macroeconomy. However, many arguments focused on very micro physical production processes, with industry leaders claiming that substitutability of Russian gas was very close to zero. Bachmann and others (2022b) argued that this "micro" or "engineering view" of substitution is too narrow and misses important mechanisms through which the macroeconomy would adapt to an import stop.

Macro elasticities of substitution are larger than the corresponding micro elasticities. That is, even if substitution is completely impossible at the very micro level, this does not necessarily mean that there is no substitution in the aggregate economy. Technically, single production processes may be very close to displaying a zero elasticity of substitution (Leontief), but they may still aggregate up to an economy with a positive and potentially much higher elasticity of substitution. The observation that zero or low substitution at the micro level does not necessarily imply low substitution at the macro level, goes back to a classic paper by Houthakker (1955) who showed that an economy in which individual firms that have Leontief production technologies (i.e., individual elasticities of substitution of zero) can aggregate up to a Cobb-Douglas aggregate production function (i.e., an aggregate elasticity of substitution of one). More generally, it is a classic result in macroeconomic theory that the elasticity of substitution increases with the level of aggregation (Jones 2005; Oberfield and Raval 2021).

The apparent lack of substitutability is thus a classic "micro-to-macro fallacy" (of which there are a number in economics). It also provides a straightforward explanation for why many industry representatives seem

13. Sinn writes: "Many have called for an embargo on European imports of Russian gas, arguing that this would [come] at minimal cost to Europe in terms of lost GDP [including a hyperlink to Bachmann and others (2022b)]. A new study exposes this argument for the fantasy that it is. . . . Due to the terms-of-trade effect, the welfare of consumers of gas and gas-intensive goods would decline as the price of these now-imported items increases [an effect missed by considering real GDP]" (Sinn 2022, par. 2–4). That $GNE = C + I + G$ is not subject to this criticism is easiest to see in models without investment or a government in which it just equals welfare-relevant consumption C. A possible reason for Sinn's misguided criticism is that he did not read Bachmann and others (2022b) past the executive summary, thus missing the analysis in terms of GNE.

to believe that the world is one of little substitution (a "Leontief world"): they are actually right at the micro-micro level, and this "engineering viewpoint" biases them to also view the macroeconomy in this fashion. (Of course, the alternative explanation for the apparent belief is simply industrial lobbying, a point we return to later.)

II.C. The Importance of Time: The Le Chatelier Principle and Seasonality of Gas Demand

Another important observation about elasticities of substitution is that they increase with the time horizon over which the substitution ought to take place. Switching a glass melting furnace from gas to fuel oil from one day to the next is probably impossible, but given enough time, such a switch may well be feasible.[14] The idea that elasticities increase with time has become known as the Le Chatelier principle (Samuelson 1947; Milgrom and Roberts 1996).[15] It is also well known that gas demand is strongly seasonal, with demand being about three times higher in winter than in summer, primarily due to households using gas for heating.[16]

The Le Chatelier principle in combination with the seasonality of gas demand was one important reason why Bachmann and others (2022b) argued that an immediate, cold turkey import stop in April 2022 would not entail much larger economic costs than an import stop in the summer or early fall. Because a cutoff at the beginning of April would have coincided with the end of the previous heating period and a drop-off in household demand, gas supplies would have been sufficient at any point in time to satisfy both industrial and household gas demand and to avoid shortages.

In particular, also in the case of an April 2022 import stop, industry would have had time until the following winter to conserve and substitute gas. While a cold turkey import stop would have resulted in less gas imports from Russia and thus a larger required demand reduction, it would have arguably also sent the signal to industry to start substituting and adapting at full speed already from April rather than only later in the summer and thus longer adjustment times until the next winter (i.e., larger elasticities of substitution by the Le Chatelier principle). See section IV for a detailed analysis of the importance of gas imports from Russia from April to August 2022.

14. Switching glass melting furnaces from gas to fuel oil is not a hypothetical example but actually happened; see example 13 in the collection of thirty-six substitution examples in online appendix E.

15. Atkeson and Kehoe (1999) build models of energy use that rationalize the Le Chatelier principle.

16. See, for example, figure 2 in Bachmann and others (2022a).

II.D. Modeling Supply Chains and International Trade: Cascading Effects and Substitution via Imports

Much of the German debate in February and March 2022 centered around cascading effects in production, the idea that a cutoff from Russian gas would not only affect energy-intensive upstream sectors but also subsequently take down the entire supply chain and industrial sector with it. For example, a drop in gas supply would lead to a drop in glass production (a very gas-intensive product), which would lead to a drop in the production of bottles, then a drop in the production of medicine, which would affect the ability to provide hospital care, and so on. Theoretically, if production were Leontief and elasticities of substitution were zero everywhere along the supply chain, then a 20 percent drop in gas supplies would lead to a 20 percent drop in glass production, the production of bottles, and so on, and ultimately to a 20 percent drop in economy-wide industrial production.

To take the possibility of knock-on effects along the supply chain seriously, Bachmann and others (2022b) modeled such supply chains using the Baqaee and Farhi (2024) model. The Baqaee-Farhi model is a multi-sector model with rich input-output linkages and in which energy is a critical input in production. The model is designed to address questions in which supply chains or production networks play a key role, specifically how a shock to an upstream product (e.g., an energy input) propagates downstream along the supply chain, that is, the cascading effects discussed above. The model features forty countries as well as a composite country representing the rest of the world, and thirty sectors with interlinkages that are disciplined with empirical input-output matrices from the World Input-Output Database (Timmer and others 2015). Each entry of the World Input-Output matrix represents a country-sector pair; for example, we use data on the expenditure of the German "Chemicals and Chemical Products" sector on "Electricity, Gas and Water Supply" and how much of this expenditure goes to different countries, say how much goes to Germany itself and how much to Russia. The model features a nested CES structure.

The idea that input-output linkages can serve as a propagation mechanism for such shocks is well established in the literature. See Carvalho and Tahbaz-Salehi (2019) for a review of this literature and Carvalho and others (2021) for a prominent example studying the propagation of the 2011 Japan earthquake that destroyed the Fukushima nuclear plant.

As just mentioned, the Baqaee-Farhi model features not only multiple sectors but also multiple countries and thus international trade. The analysis

using this type of model points to one margin of substitution that turned out to be important in practice: substitution of gas-intensive products via imports. Intuitively, it is not necessary for German producers to substitute gas itself; instead, they can substitute the energy-intensive inputs they use in production, like ammonia, and they can do so via trade by importing those goods from another country. In this way, producers effectively import gas "embodied in" these inputs. Of course, this type of substitution via imports comes with some loss in production in the importing country (in this case, Germany). However, these losses may be small, and on the flip side, this substitution stops the notorious cascading effects.

Finally, it is worth noting that an empirically disciplined multi-sector model like the Baqaee-Farhi model reflects an important feature of modern advanced economies: manufacturing typically accounts for a moderate share of aggregate economic activity. This is true even for Germany, which is often viewed as an industrial powerhouse: German manufacturing accounts for only about 23 percent of total employment and 25 percent of value added.[17] This is a natural consequence of the structural transformation process during which manufacturing activity is replaced by the service sector. Put differently, some observers seem to be under the mistaken impression that the structure of the German economy is still that of earlier time periods like the 1970s, during which energy shocks had large negative effects.

II.E. A Useful Tool: The Baqaee-Farhi Sufficient Statistics Approach

In a number of papers, Baqaee and Farhi have popularized the use of second-order approximations to obtain analytical results in complex multi-sector models. Bachmann and others (2022b) use a variant of this approach to obtain a useful sufficient statistics formula that allows for quick back-of-the-envelope calculations.

The key idea of the approach is that the extent to which the upstream energy supply shock propagates through the production chain shows up in a sufficient statistic, namely, the change of the energy expenditure share in GNE induced by an import stop. Intuitively, when there are important bottlenecks along the supply chain and elasticities of substitution are low, energy prices skyrocket when energy supply falls, which implies that the energy expenditure share rises strongly.

It is relatively easy to verify that this insight is correct in the context of the simple aggregate production function (see online appendix A). Perhaps

17. See the appendix in Bachmann and others (2022b), which documents these numbers using Eurostat data.

surprisingly, Bachmann and others (2022b) show that it is also true in the much more complex multi-sector environment of Baqaee and Farhi (2024). Denoting gas imports by m_G and their price by p_G so that the gas expenditure share in GNE is given by $p_G m_G/GNE$, the effect of a shock to gas imports $\Delta \log m_G$ approximately equals

$$\Delta \log GNE \approx \frac{p_G m_G}{GNE} \times \Delta \log m_G + \frac{1}{2} \times \Delta\left(\frac{p_G m_G}{GNE}\right) \times \Delta \log m_G. \tag{2}$$

The intuition for the second term is the one we already discussed: the change in the GNE share of gas imports $\Delta\left(\frac{p_G m_G}{GNE}\right)$ summarizes in a succinct fashion the substitutability implied by model choices about elasticities, the input-output structure, and so on.

The formula can be used for back-of-the-envelope calculations as follows. Consider, for example, a drop in gas imports by 30 percent so that $\Delta \log m_G = \log(0.7)$. The share of gas expenditure in GNE $\frac{p_G m_G}{GNE}$ equals about 1.2 percent. The second-order approximation also requires a number for the change in the expenditure share $\Delta\left(\frac{p_G m_G}{GNE}\right)$, a number that was not yet available in the data at the time of writing by Bachmann and others (2022b). In one of their calculations, they assumed that this share would quadruple to 4.8 percent. Using these numbers, the GNE losses are given by

$$\begin{aligned} \Delta \log GNE &\approx 1.2\% \times \log\left(0.7\right) + \frac{1}{2} \times \left(4.8\% - 1.2\%\right) \times \log\left(0.7\right) \\ &\approx -1\%. \end{aligned} \tag{3}$$

More generally, formula (2) can be used to bound the GNE loss from the shock: above a certain GNE loss number, the strong complementarities and cascading effects required to get there would imply an unreasonably large increase in the gas expenditure share, say, to 20 percent of GNE. It is worth noting that this logic applies not just to the Baqaee-Farhi model but also to a much wider class of general equilibrium models. Other analyses of import supply shocks should therefore always examine the model's predictions for changes in expenditure shares for their reasonableness.[18]

18. See also Berger and others (2022), who put the sufficient statistics approach based on formula (2) to good use.

II.F. Additional Arguments and Omissions from the Analysis

Less than two weeks after the release of Bachmann and others (2022b), we added a detailed appendix to the paper with a number of historical real-world examples that show how firms and households have found ways to substitute in adversity.[19] These include the Chinese rare earths embargo against Japan, the shutdown of the Druzhba pipeline, and various examples from World Wars I and II. There is one particularly relevant case study we were not aware of at the time, namely, the case of Chile getting cut off from Argentinean gas in 2007—see the illuminating discussion by Velasco and Tokman (2022) who were the Chilean finance and energy ministers at the time.

As the "what if" paper was clear to emphasize, our analysis used a real model with no further business cycle amplification and therefore omitted some of the channels through which a large energy supply shock may affect the economy. In particular, our model omitted standard Keynesian demand-side effects in the presence of nominal rigidities as well as amplification effects due to financial frictions. To be clear, our flexible-price model did include what many lay people would call demand-side effects, namely, that skyrocketing relative prices of energy erode purchasing power and consumer welfare. But it omitted the feedback from the drop in aggregate consumption to production and employment that is operational in Keynesian models with nominal rigidities and high marginal propensities to consume. To acknowledge such missing mechanisms, we added a "safety margin" to the results of their model simulations. In particular, our largest number in the "what if" paper was a GNE loss of 2.3 percent (see table 2 in the paper) which we rounded up to 3 percent when presenting our headline numbers (see the abstract). Perhaps reassuringly, work by our coauthor Christian Bayer (Bayer, Kriwoluzky, and Seyrich 2022), published a few weeks after the "what if" paper, as well as Pieroni (2023) used quantitative Heterogeneous Agent New Keynesian (HANK) models to take into account such Keynesian multiplier effects and largely confirmed our original results.[20]

19. See "Supplement to 'What If? . . .': Real-World Examples of Substitution and Substitution in the Macroeconomy" available at https://benjaminmoll.com/RussianGas_Substitution/.

20. Bayer, Kriwoluzky, and Seyrich (2022) and Pieroni (2023) modeled exactly the same gas supply shock as we did in Bachmann and others (2022b) but in HANK models. Bayer, Kriwoluzky, and Seyrich (2022) found that the upper bound of economic costs stayed below 3 percent of GDP, that is, below the "safety margin" we left ourselves, whereas Pieroni (2023) found that economic costs could reach up to 3.4 percent, that is, just outside our upper bound.

The main reason for these omissions was not that we deemed these effects unimportant. Instead, it was simply that we wrote the "what if" paper in a rush (ten days) and therefore, given time constraints, had to make choices about what channels to include in our analysis and what to leave out. We will revisit these points in section III.F, where we discuss which of these omissions were important with the benefit of hindsight and lessons for future analyses of similar scenarios.

III. How the Adjustment Happened: Adaptation and Substitution by German Industry and Households

A year after the final cutoff from Russian gas, we can take stock of what happened to the German economy. The most recent GDP numbers for the German economy were published at the end of July 2023. Prima facie, the evidence seems to support the original argument of the "what if" paper. Germany was partially cut off from Russian gas in June 2022 and completely cut off in August 2022, but the country did not go into a deep depression. Moreover, as shown in figure 1, German GDP not only did not collapse, but actually expanded by close to 2 percent for the entire year 2022. Even during the peak of the heating season of the 2022–2023 winter, Germany only experienced a mild one-quarter contraction, with GDP falling by 0.4 percent in the fourth quarter of 2022 and stagnating at close to 0 percent GDP growth during the first three quarters of 2023.[21]

Using the empirical evidence now at hand, this section documents how the adjustment actually played out. As we see now in greater detail in the rearview mirror, the economy showed a tremendous ability to adapt that was widely underestimated. Producers partly switched to other fuels and imported products with high gas content, while households adjusted their consumption patterns. Overall industrial production decoupled from production in energy-intensive sectors (which did see large drops) and was hardly affected. To lend some color to the statistics of this section, online appendix E collects thirty-six concrete cases of substitution and adaptation that show how German firms and households weaned themselves off Russian gas.

21. Other European countries also withstood Russia's weaponization of natural gas remarkably well. According to the most recent Eurostat GDP flash estimates for 2023:Q2 (Eurostat 2023), both the European Union and the euro area expanded in the first two quarters of 2023, and only a handful of individual member countries like Czechia and Estonia have experienced (shallow) recessions (defined as two consecutive quarters of negative GDP growth) since the beginning of 2022. The exception is Hungary, which has seen four consecutive quarters of negative GDP growth since 2022:Q3.

III.A. Germany's Changing Gas Balance: Large Adjustments on Both the Demand and Supply Sides

The end of Russian gas imports left a large gap in German gas supplies. How did the country adjust to close this gap? Was the adjustment primarily on the demand side, that is, lower gas consumption, or supply side, that is, increased imports from third countries? Figure 4 shows the change of the German gas balance for the period from July 2022 (when Russia cut gas supplies substantially; see section I) to March 2023 (the end of the heating period), compared to the preceding three years.

The cutoff from Russian gas reduced supply by 41 percent of total consumption in previous years.[22] This gap was filled by large adjustments on both the demand and supply sides. Additional supplies from third countries (like Norway, Algeria, and the United States) accounted for 33 percent of the gap, while gas demand in 2022–2023 was about 20 percent lower compared to the 2019–2021 average.[23] Finally, an additional 10 percent of annual consumption was used to increase storage levels, in part necessary because some storage facilities were Russian-owned and had been purposely kept empty. We postpone further discussion of the supply side to section III.E, where we break down the sources of the new gas supplies and highlight the insurance function played by European and global market integration.

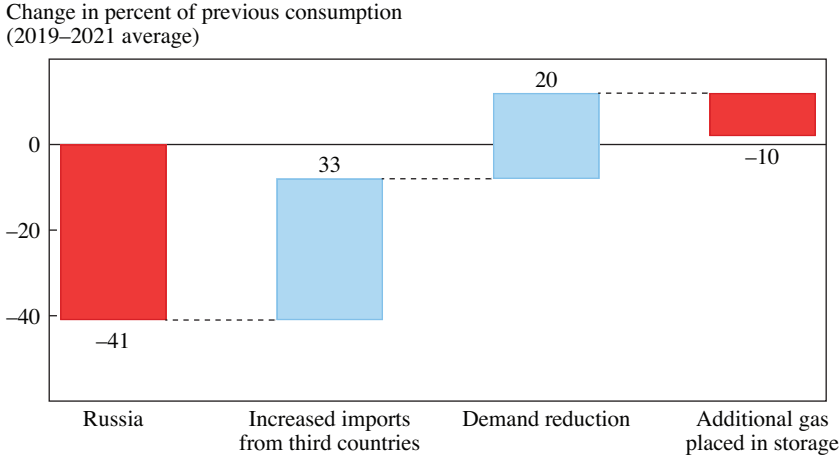
Zooming in on the demand side, table 2 breaks down the 20 percent demand reduction into its key components using data from Ben McWilliams and Georg Zachmann's European Natural Gas Demand Tracker.[24] With the exception of electricity generation, where gas demand for power generation fell only by a small single-digit amount, industrial demand fell by 26 percent and household demand by about 17 percent.

22. This number differs from the 55 percent number in table 1 for two reasons associated with time periods. First, table 1 reports Russian imports as a percentage of average consumption over the whole year, whereas figure 4 reports them as a percentage of average consumption over the nine-month period from July to March. Average gas consumption in the July to March period is higher than over the whole year because it puts a higher weight on the heating period, thus resulting in a higher denominator and lower percentage value. Second, the numerator also differs: table 1 reports Russian gas imports for 2021, whereas figure 4 computes the reduction relative to a time period ending in March 2022. These are different because Russian imports (Yamal and Ukraine transit flows) already dropped considerably in early 2022.

23. On the supply side, we take into account not only direct imports to Germany but also indirect imports via third countries as well as reexports within the European Union. For comparison, online appendix figure B.2 plots the direct flows.

24. Bruegel, "European Natural Gas Demand Tracker," https://www.bruegel.org/dataset/european-natural-gas-demand-tracker.

Figure 4. Germany's Changing Natural Gas Balance



Source: Eurostat (database code nrg_ti_gasm); Ben McWilliams and Georg Zachmann's European Natural Gas Demand Tracker; and Aggregated Gas Storage Inventory (AGSI).

Note: The figure compares German natural gas imports, consumption, and storage change for the period from July 2022 to March 2023 to the corresponding average from 2019 to 2021. On the supply side, we take into account not only direct imports to Germany but also indirect imports via third countries as well as reexports within Europe. More details, including sources, are in online appendix B.

Table 2. Large Demand Reduction by Industry and Households

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
	2022–2023 consumption (TWh)	*Baseline consumption (TWh)*	*Reduction relative to baseline (TWh)*	*Percentage reduction*	*Hypothetical adjustment (percent)*
Total	642	799	157	20	25
Industry	276	373	98	26	26
Households	281	339	58	17	16
Power	85	87	1	2	45

Source: European Natural Gas Demand Tracker; and Bachmann and others (2022b).

Note: The table summarizes gas consumption over the period July 2022 to March 2023 (column 1) and compares it to average consumption in the same months in the years 2019 to 2021 (column 2). Column 5 refers to predictions about a hypothetical adjustment path made in Bachmann and others (2022b) in early August 2022, ahead of the gas cutoff. The data source provides a more detailed methodology for the calculation of demand, but the key assumptions are as follows: gas consumption is measured separately for so-called RLM meters (large consumers directly connected to the transmission grid) and SLP meters (small consumers). "Households" refers to small consumers (SLP) and therefore also includes commerce and small businesses. "Power" refers to gas used in electricity generation, which we calculate from power output of gas-fired power plants and assuming a plant efficiency of 50 percent. Consumption by industry is calculated by removing gas used for power-generation from RLM consumption. That the numbers in the last row seemingly do not add up is due to rounding.

These numbers are not far off the adjustment path described in our second paper ahead of the gas cutoff (Bachmann and others 2022a), in which we counted on a 26 percent demand reduction by industry and 16 percent by households. However, we substantially overestimated the potential for gas savings in electricity generation. As we will discuss later, this had a lot to do with specific elements of bad luck in electricity generation (the shortfall in French nuclear energy production and the drought in Europe, which reduced available hydropower substantially). The demand reduction was supported by good incentives for savings for households emanating from the proposals of an expert commission, as we will discuss below.

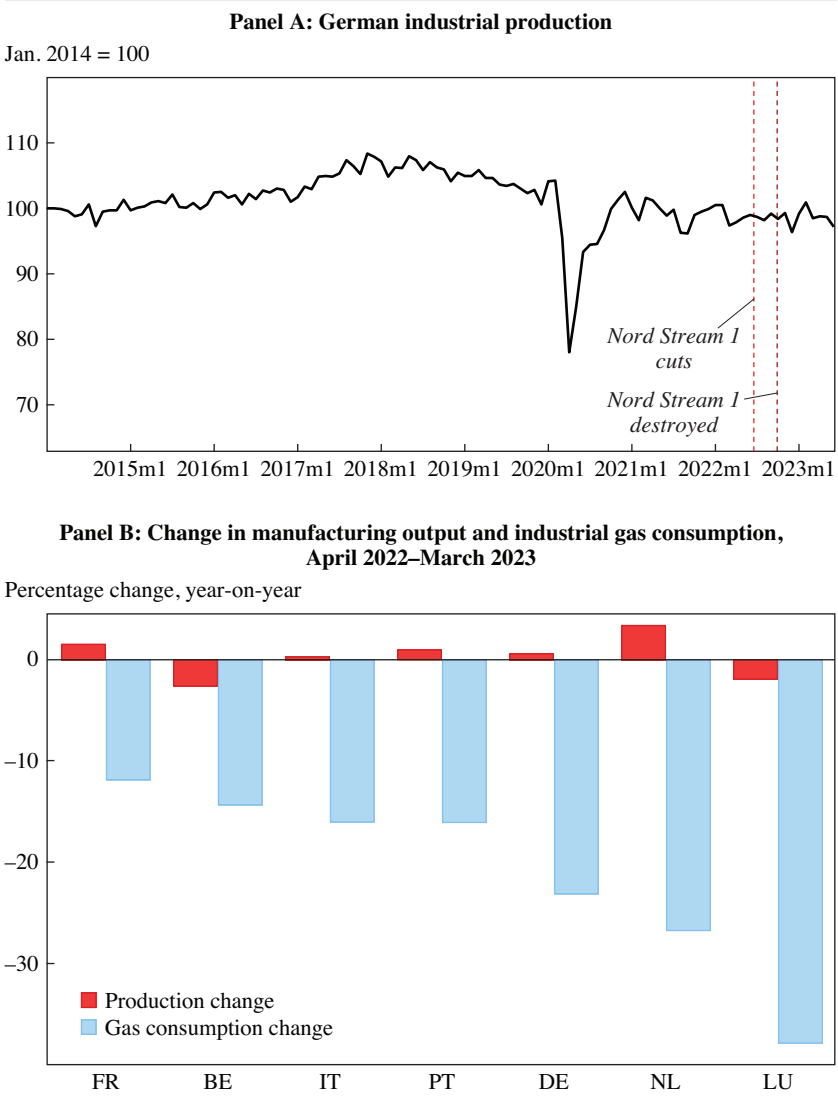
Section II.E emphasized a key sufficient statistic, the change in Germany's gas expenditure share. While our original analysis was forced to speculate about the future evolution of this statistic, online appendix figure B.3 plots this expenditure share using the evidence now at hand. Before the 2021–2022 winter, natural gas accounted for around 1 percent of Germany's total expenditure (GNE). As Russia weaponized and restricted gas supplies, skyrocketing prices meant that this expenditure share increased sharply to around 4 percent of GNE. This quadrupling of the gas expenditure share turned out to be in line with the experiment we described in section II.E and for which the Baqaee-Farhi sufficient statistics approach predicted a 1 percent GNE loss.

III.B. Industry

Taking a closer look at the 20 percent aggregate demand reduction over the past heating period, the evolution of gas consumption and output in the industrial sector is of particular interest as much of the original arguments on the effects of the cutoff focused on the short-run substitutability of gas in industrial production. We already know that, in the aggregate, industrial gas usage decreased by 26 percent relative to previous years (table 2). Importantly, this sharp reduction in gas usage was not accompanied by large output drops, as many had feared.

Figure 5 plots industrial production and gas consumption in Germany and six other European countries. As a benchmark, recall from section II the key prediction that a Leontief zero-substitutability production structure implies that production falls one for one with gas consumption. That is, if elasticities of substitution in industry had been truly zero, Germany should have seen overall industrial production fall by around 26 percent, as the drop in industrial gas usage would have cascaded through the entire supply chain. Figure 5 demonstrates that not only in Germany, but also across the rest of Europe, industrial production looks nothing like this Leontief case.

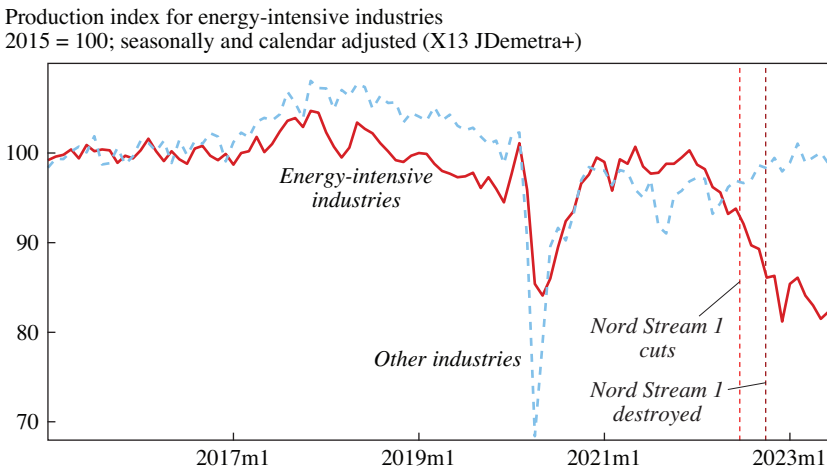
Figure 5. Industrial Production in Germany and Europe Looks Nothing Like Leontief



Source: Destatis; European Natural Gas Demand Tracker; and Eurostat.

Note: The industrial production data in panel A are from table 42153-0001 of the German economic sectors statistics, available through the German statistical agency, Destatis, at https://www-genesis.destatis.de/. The index is normalized to 100 in January 2014. Panel B compiles gas demand data for industries from Ben McWilliams and Georg Zachmann's European Natural Gas Demand Tracker, with industrial output data from Eurostat (database code: sts_inpr_m).

Figure 6. Decoupling of Overall Industrial Production from Energy-Intensive Sectors



Source: Destatis; and Vogel, Neumann, and Linz (2023).

Note: Data are from Destatis, figure 5, "Bedeutung der energieintensiven Industriezweige in Deutschland" [Importance of energy-intensive industries in Germany]. Energy-intensive industries are: (1) paper and paper products, (2) coke and refined petroleum products, (3) chemicals and chemical products, (4) basic metals, and (5) other nonmetallic mineral products, which together account for a total of 16.4 percent of overall industrial production in the base year 2015 (Vogel, Neumann, and Linz 2023). The index for overall industrial production is a weighted average for energy-intensive industries and for other industries with weights 16.4 percent and 83.6 percent. This allows us to back out the index for other industries from the index for overall industrial production and that for energy-intensive industries.

In Germany, industrial production did not fall meaningfully and even rose compared to the previous year, depending on the month of comparison. On the European level, hardly any correlation can be observed between reductions in gas consumption and manufacturing output. In the Netherlands, for instance, gas consumption fell by almost 30 percent while industrial output overall increased significantly.

We next ask what sectors were most affected by the gas cutoff, and whether and to what extent there were knock-on effects along the supply chain. Unfortunately, the German statistical agency, Destatis, would only release detailed data for 2022 gas usage by industry sector in October 2023. However, we can use preexisting classifications of industries into more and less energy-intensive sectors to gain a better understanding of the actual adjustment processes.

We find clear indications that production in energy-intensive sectors was strongly affected. Figure 6 displays the time path for production in energy-intensive industries using the classification of Destatis alongside

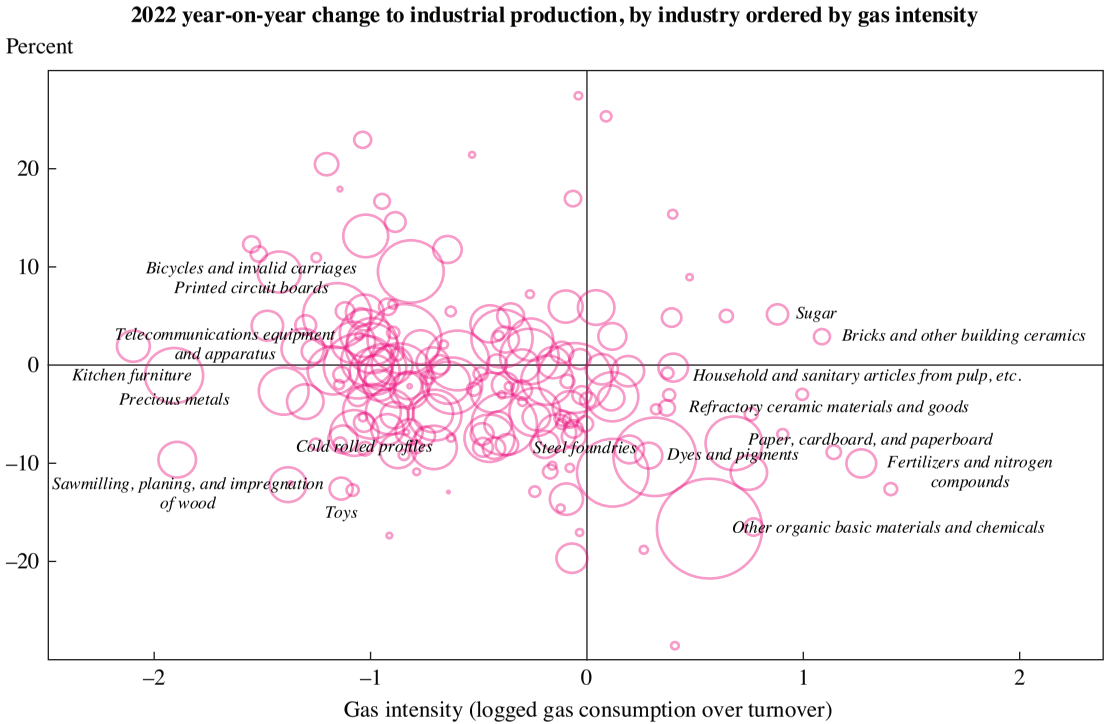
production in other industries. As can be seen from the graph, production in energy-intensive sectors dropped by close to 20 percent since gas prices started skyrocketing in early 2022.[25] However, industrial production of other sectors declined only slightly. Importantly, this observed decoupling between energy-intensive production and production of other sectors is the polar opposite of the much-feared cascading effects discussed earlier. Figure 6 (along with the results in figures 7 and 8 below) shows that in an open economy with substitution possibilities, sharp declines in output in some upstream sectors do not necessarily lead to large contractions in downstream industries. At each point in the production network substitution possibilities exist.

Figure 7 conducts a more granular analysis using our own measure of gas intensity at the sectoral level, with gas intensity defined as an industry's past gas consumption relative to its turnover. As expected, there is a clear negative correlation between changes in industrial production and gas intensity, with the most gas-intensive sectors seeing the largest drops in industrial production. However, not just the slope of the relationship is interesting but also the level. In particular, while energy-intensive sectors like chemicals, paper, and fertilizer did see sharp drops in production (presumably because they also saw substantial drops in gas consumption), many other sectors saw no drops or even increases in production. Instead, in a "cascading-effects view" of the world, industrial production should have fallen in all sectors regardless of how energy intensive they are, because the initial negative gas supply shock to gas-intensive sectors should have taken down the entire supply chain. Figure 7 thus again shows no evidence of cascading effects and instead shows more of the decoupling already evident in figure 6.

When Destatis releases 2022 gas usage by industrial sector in October 2023, it would be interesting to correlate the drops in industrial production in figure 7 with the drops in gas usage. Such a sectoral version of figure 5 (panel B) would provide the sharpest test of the extent of substitution along the supply chain by answering the question: whether production only fell in particular gas-intensive sectors with large drops in gas usage;

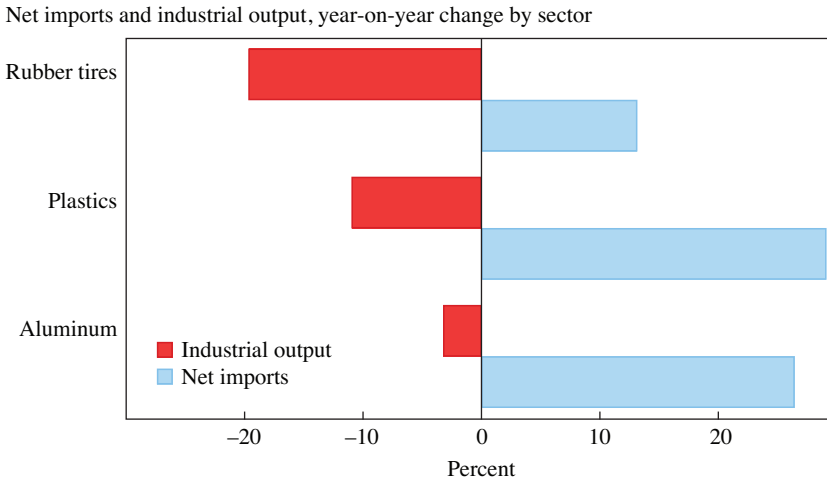
25. An interesting question is how close this large production drop in energy-intensive sectors was to the Leontief benchmark of a one for one drop with gas consumption. Since data on gas usage by sector have not been released at the time of writing, we cannot answer this question in this paper. A natural conjecture is that the gas usage in these sectors dropped by more than the 26 percent reduction for industry as a whole, which would imply that not even production in those sectors behaved like in the Leontief case.

Figure 7. Sectoral Output Change and Energy-Intensity of Industrial Sectors



Source: Destatis.
Note: Industrial production and energy consumption data are merged according to Klassifikation der Wirtschaftszweige (WZ) sector codes.

Figure 8. Illustrative Examples of Substitution via Imports



Source: Destatis; and Eurostat.

Note: Destatis industry-level data for industrial production are mapped to trade data from Eurostat (database code DS-045409). For rubber tires ("New Pneumatic Tyres, of Rubber") the WZ code for the classification of economic sector by Destatis is 2211 and the Harmonized System (HS) code for global product classification is 4011. For plastics ("Plastics and Articles Thereof") the WZ code is 2016 and the HS code is 39. For aluminum ("Aluminium and Articles Thereof") the WZ code is 2442 and the HS code is 76.

or whether these production drops cascaded further downstream and even affected sectors that do not consume any gas or experienced no drops in gas usage.

Figure 8 provides some illustrative examples for the substitution via imports emphasized in section II.D by plotting output change and import growth for a number of selected energy-intensive industrial sectors like rubber, plastics, and aluminum production. We observe substantial increases in net imports of energy-intensive products. While the correlation with the reduction of output on the industry level is less close, substitution via imports was likely an important channel through which gas savings could be realized with small effects on the overall economy.

A study by Mertens and Müller (2022) provides additional support for the hypothesis that substitution via imports was likely important in practice. Using a more fine-grained product-level analysis, they show that only three hundred specific products account for about 90 percent of industrial gas consumption in Germany. They then argue that these products are heavily traded on the world market and therefore likely more easily substitutable via imports.

As already noted, online appendix E collects thirty-six concrete cases of substitution of gas and gas-intensive products by German firms and households. One of these is worth restating here because it illustrates well the substitution via imports just discussed. When gas prices skyrocketed in Germany and Europe, chemicals giant BASF drastically reduced the production of ammonia (a very gas-intensive product) at its Ludwigshafen site. BASF then switched to producing ammonia in its other plants around the world including in the United States where gas prices were much lower, and more generally, to importing ammonia from other countries. A newspaper article noted that "this substitution via the world market [is] relatively easy" (Höltschi 2022, par. 12).[26] What is worth noting here is that substitution via imports can sometimes even happen *within* the same firm. It is also worth contrasting BASF's apparent substitution prowess with its chief executive's statement about the destruction of the entire economy quoted at the beginning of our paper.

Finally, there is some high-level and suggestive evidence that lower industrial gas demand was, at least in part, due to skyrocketing gas prices—see Ruhnau and others (2023), in particular the downward-sloping time series relationship between monthly prices and quantities in their figure 5(b). The endogeneity of both prices and quantities as well as the complexity of the gas market, mean that this evidence should not be interpreted as causal. But it is nevertheless worth highlighting that high prices were associated with reductions in industrial gas demand.

III.C. Households

Consumption by households and other small consumers represents around 42 percent of overall gas consumption.[27] Because households use gas overwhelmingly for heating, their demand is both highly seasonal and influenced by weather variations (see section IV). Overall, German households consumed 17 percent less gas in the period from July 2022 to March 2023 than in the same period in the three preceding years (table 2).

Online appendix figure B.4 shows that demand reduction by households was significant even when controlling for temperature. While temperature-controlled household demand in January and February 2022 was above

26. See also cases 2 and 15 in online appendix E.

27. As explained in the note to table 2, what we term household gas consumption is consumption by SLP consumers (small consumers not directly connected to the transmission grid), and therefore includes not just households but also some commerce and small businesses.

average, from March 2022, that is, after the war started, it increasingly fell below average. This indicates that households actively reduced their gas consumption. A lot of this saving might have been behavioral, that is, reducing room temperature or heating fewer rooms. But over time we might see more and more structural savings based on investments, ranging from light-touch investments in insulating drafty doors and windows to substantial capital spending on replacing gas boilers with heat pumps.

Disentangling the causes of these quite significant household gas demand reductions will provide important lessons for policymakers and the energy industry. The early demand reductions in March 2022, when high wholesale prices had not yet translated into increasing retail prices, indicate that the shock of the crisis, discussions about emptying gas storages, and public appeals had some effect on household behavior. There was, however, only a very limited federal level gas saving campaign. It had a budget of only 40 million euros—that is, about 50 cents per German citizen—and was targeted at energy switching not at energy saving, and it was not evaluated.[28] This was maybe over worries that a hard savings campaign would rather upset the population (Deutscher Bundestag 2022). More importantly, there was no federal public program to support demand-side investments into gas savings, while at the same time billions were spent on the supply side. On the regional, state, and local levels, campaigns have been run by administrations and gas suppliers.

In general, German retail prices are sticky and billing often happens only once a year. Assessing the impact of retail prices on household gas consumption is held back by a lack of public granular data and has only just begun. Such granular data will be key, as households' exposure to rising gas prices differed widely depending on the region they lived in, their gas suppliers, their gas consumption patterns, and most importantly the supply contracts they were on. As the wholesale price explosion was passed through differently to different customers, the demand reduction patterns might also differ.

Still, over time an increasing share of consumers saw their gas prices go up significantly. All new and renewed retail gas contracts since March 2022 featured significantly higher prices so that more and more consumers were affected by increasing prices over time. By autumn of 2022, a substantial share of consumers had been confronted with drastically increased prices. This visibly impacted demand. Gas prices across countries and

28. The campaign was called "Energiewechsel," which means "energy switch."

changes in gas prices correlate with gas demand reductions during the crisis (McWilliams and others 2022). That is, countries with the highest increase in household gas prices saw the strongest reduction in gas demand in the European Union.

This also shifted the political dynamics for the state to intervene. In September, the federal government set up an expert commission to discuss sensible policies to help consumers without increasing demand (see section III.D), while at the same time temporarily reducing value-added tax for natural gas from 19 percent to 7 percent, muting the price signal for consumers at the expense of German taxpayers (Bundesregierung 2022a).

Analogous to the case of industrial gas demand, there is some high-level and suggestive evidence that high prices were associated with household demand reductions; see Ruhnau and others (2023), in particular the downward-sloping relationship between monthly prices and quantities in their figure 5(a), though with the same caveats as in the case of industrial gas demand (see the discussion above).

III.D. Policy Choices Matter: Germany's Alternative to a Price Cap

Skyrocketing gas prices in the summer and fall of 2022 put substantial strains on the finances of both households and firms, leading to calls for policy intervention to support households and firms. In contrast to policymakers in many other European countries, German policymakers refrained from imposing a price cap on natural gas and instead opted for lump-sum transfers based on households' and firms' historical gas consumption. We briefly review this scheme here for two reasons. First, the scheme is interesting from an economic perspective in that it provides relief by aiming to target the income effect of higher gas prices while leaving substitution effects intact, akin to what Mas-Colell, Whinston, and Green (1995) term "Slutsky compensation." Second, the scheme is an interesting blueprint for future government interventions to alleviate the hardship in the face of rising commodity prices.

The policy was based on the proposal of a commission composed of various stakeholders (such as union and industry leaders) as well as a number of economists, including our coauthors Christian Bayer and Karen Pittel (ExpertInnen-Kommission Gas und Wärme 2022). Precursors of this scheme were proposed by Bayer in Bachmann and others (2022a, 2022b). As has been widely discussed, the official name of the German policy scheme, which translates as "gas price brake," is a misnomer, and "gas cost brake" may instead have been a more accurate name. This is because the scheme caps a household's or firm's total expenditure rather than the

marginal price of an extra kWh of gas, which remains equal to the pre-intervention market price.[29]

Figure 9, panel A, graphically illustrates the German scheme using a numerical example. The *x*-axis plots a household's current gas consumption as a percentage of its historical consumption, which is assumed to be 10,000 kWh. The *y*-axis plots the household's gas bill in euros as a function of its gas consumption under a number of scenarios of gas prices and policy interventions. Initially, the gas price paid by households is at 5 cents per kWh, resulting in a gas bill of 500 euros (dash-dotted line). Now gas prices skyrocket by a factor of 5 to 25 cents per kWh so that the gas bill of a household consuming 10,000 kWh of gas is not 500 euros but 2,500 (solid line with circle markers). What are the effects of various policies to support households? One option is a price cap, say at 12 cents per kWh (dashed line). As desired, this brings down the gas bill from 2,500 to 1,200 euros. But it also comes with a problem: it strongly reduces the household's incentive to reduce gas consumption relative to the high price (the dashed line is flatter than the solid line with circle markers).

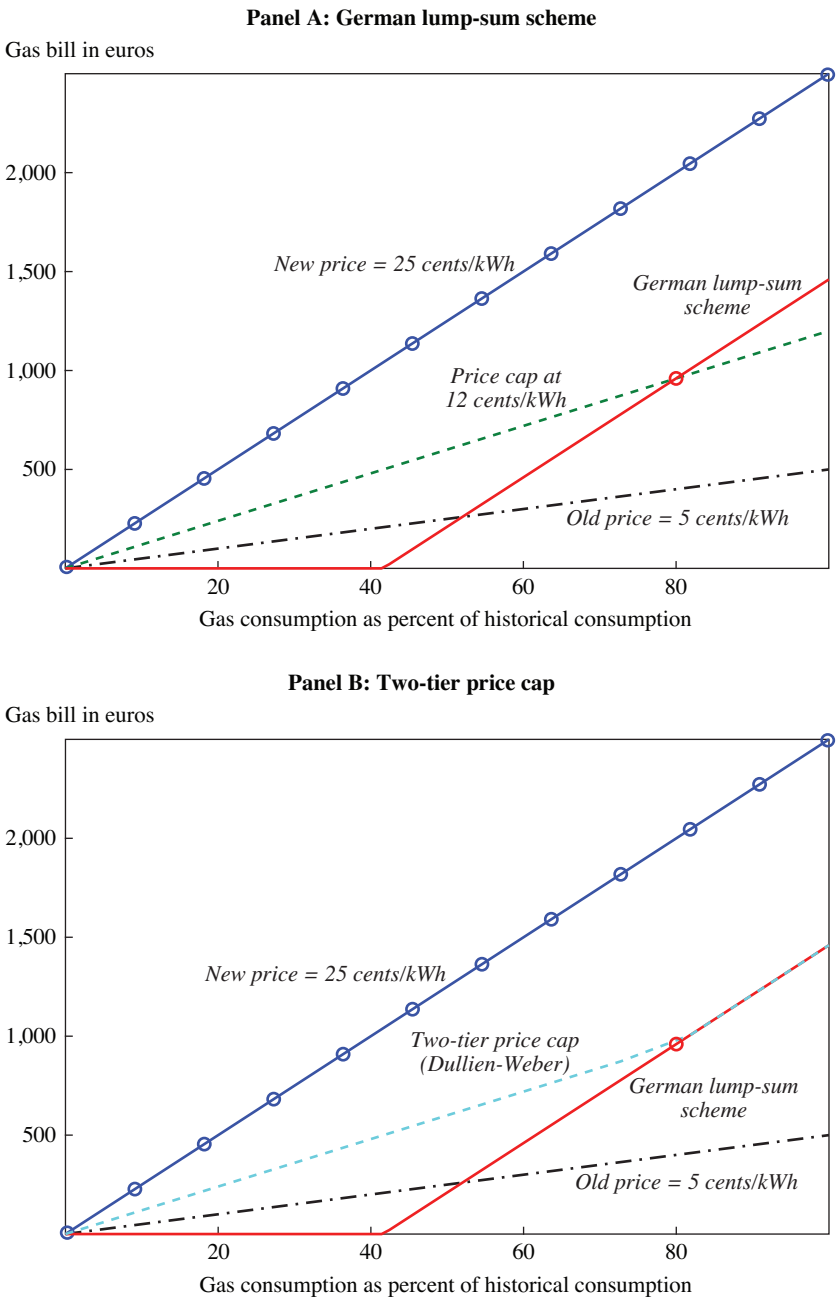
The German policy is represented by the solid line. Households receive a transfer (credit on their gas bill) equal to 80 percent of their historical consumption times the difference between the current market price of 12 cents per kWh (an estimated long-run "new normal" gas price).[30] The key observation is that, in contrast to a price cap, this transfer is not directly tied to current gas consumption (i.e., it is a lump-sum transfer) and thus preserves incentives for reducing gas consumption. Graphically the solid line has the same slope as the solid line with circle markers (though it is everywhere below the latter).[31] By using a household's historical gas consumption as the basis for calculating the size of the transfer, the scheme is nevertheless targeted toward more affected households. Skyrocketing gas prices have both an income and a substitution effect. The income effect is undesirable because it makes households poorer; in contrast, the substitution effect is desirable because it reduces gas consumption. An appealing feature of the German scheme is that it leaves the substitution effect unaffected while

29. See Bayer and others (2023) and Bundesregierung (2023) for summaries and preliminary evaluations of the scheme.

30. Bundesregierung (2022b). The transfer is capped at the total bill amount, that is, it is not possible to make money. Graphically the solid line equals zero when gas consumption drops below about 40 percent of historical consumption.

31. Of course, there *is* a relation between the German scheme (solid line) and a price cap at 12 cents per kWh (dashed line): the point where the two lines cross is exactly 80 percent of past consumption. So the dashed line for the price cap determines how much the solid line is shifted down.

Figure 9. The German "Gas Price Brake" Was a Lump-Sum Transfer and Not a Price Cap



Source: Authors' calculations.

alleviating the negative income effect. The scheme is thus a variant of what the literature has termed "Slutsky compensation" (Mas-Colell, Whinston, and Green 1995). An important point is that the German scheme is not a two-tier price cap, for example, a price cap for 80 percent of past consumption with market prices kicking in for consumption above 80 percent, as proposed by some economists.[32]

Figure 9, panel B, contrasts the two schemes graphically, with the solid line plotting the German scheme (as in panel A) and the dashed line plotting a two-tier price cap with a price cap of 12 cents per kWh for up to 80 percent of past consumption. The key observation is that the schemes differ for any consumption level below 80 percent of past consumption: while the German scheme preserves saving incentives for those who can save more than 20 percent relative to their past consumption, a two-tier price cap reduces these incentives by capping the price faced by consumers. Importantly, households reducing gas consumption by more than 20 percent turned out to be not just an academic curiosity: instead, during the 2022–2023 winter, larger demand reductions were routinely observed.[33]

III.E. New Gas Supplies and the Insurance Value of European Integration

As shown in figure 4, additional supplies of non-Russian gas to Germany played an important role in getting Germany through the 2022–2023 winter, with these imports increasing by around 33 percent relative to previous consumption. This section breaks down these imports further and highlights two main channels. First, additional gas imports into Europe made their way to Germany via the integrated European pipeline network. Second, demand reduction elsewhere in Europe freed up gas supplies that then ended up in Germany. Both channels underscore the insurance benefits of global and European market integration (Caselli and others 2019).

Considering Europe as a whole, gas imports increased significantly, with most of this increase coming from LNG, which increased by 470 TWh in the period after the Nord Stream cuts (July 2022 to March 2023), compared

32. See, for example, Dullien and Weber (2022).

33. While the average household demand reduction over the entire 2022–2023 winter was less than 20 percent (see table 2), demand reductions in particular weeks were considerably above 20 percent and often up to 40 percent. See Bundesnetzagentur, "Gasverbrauch Haushalts- und Gewerbekunden, wöchentlicher Mittelwert" [Gas consumption households and businesses weekly], https://www.bundesnetzagentur.de/DE/Gasversorgung/aktuelle_gasversorgung/_svg/GasverbrauchSLP_woechentlich/Gasverbrauch_SLP_W_2023.html. The same is presumably true for particular households or geographic areas.

to the 2019–2021 average and a more moderate contribution from pipeline imports, which increased by 110 TWh.[34] An important feature of the additional LNG imports was that they came at extremely high prices. Because global production capacities as well as the infrastructure for transporting LNG were constrained, LNG destined for other markets had to be rerouted to Europe by offering extremely high prices for individual cargoes. The small increase in pipeline imports to Europe was similarly due to the fact that production and transportation capacity could not be ramped up more quickly.

Turning to Germany individually, online appendix figure B.1 plots a version of figure 4 but with the imports from third countries broken down by ultimate source country. The largest supplier of additional non-Russian gas was Norway, contributing additional imports worth around 16 percent of previous consumption, that is, almost half of the 33 percent overall additional supplies. LNG imports were also important, contributing a combined total across all countries of 13 percent. Note that, like figure 4, the figure takes into account not only direct imports to Germany but also indirect imports via third countries as well as reexports within the European Union. This is particularly important for LNG because Germany had rejected building any LNG import infrastructure prior to the crisis and therefore had to rely instead on LNG terminals elsewhere in Europe (e.g., in Belgium, the Netherlands, and France) for most of these imports. Immediately following the Russian invasion, Germany put in motion plans finally to build LNG terminals on its coast. These made a small contribution of gas imports worth around 3 percent of previous consumption (see online appendix figure B.2).[35] The important role of gas imports from third countries, and specifically via other European countries, highlights the insurance benefits of global and European market integration.

While imports from outside Europe were instrumental for displacing Russian gas in Germany, another crucial factor for getting Germany through

34. The series for European LNG imports includes indirect imports of LNG via the United Kingdom that were then passed by pipeline into the Netherlands and Belgium. The UK pipeline flows to the Netherlands and Belgium dramatically increased to make use of extra LNG import capacity in the United Kingdom. In Europe as a whole, 20 percent of LNG import capacity was added in 2022:Q4 and 2023:Q1. See Bruegel, "European Natural Gas Imports," https://www.bruegel.org/dataset/european-natural-gas-imports.

35. The contribution of the newly built LNG terminals may seem small to readers who are familiar with the German gas debate given these were often touted as "game changers" by politicians and the media. The reason why their contribution to getting Germany through the 2022–2023 winter was not larger is that they only came online relatively late, with the first LNG terminal (Wilhelmshaven) opening on December 17, 2022.

the 2022–2023 winter was the demand reduction elsewhere in Europe. This is because additional imports to Europe replaced only about two-thirds of Russian imports so that an additional fall in demand was needed.[36] In the European Union as a whole, gas demand declined by a substantial 18 percent or 630 TWh in the period from July 2022 to March 2023 compared to the 2019–2021 average.[37] Gas consumption fell substantially not only in countries that were highly dependent on Russia but also in others that were not. This freed up additional gas supplies for those countries most in need. A political commitment to reducing gas consumption by at least 15 percent (European Commission 2022) likely contributed to this EU-wide demand reduction, specifically because it entailed a commitment to letting markets work despite the very high prices that were adversely impacting domestic industrial and household consumers alike. In summary, high prices discouraged demand all over the European Union, high prices at the entry points into the European system drew international volumes into Europe, and intra-European gas price differentials pulled gas flows into the countries most in need of volumes to replace Russian supplies, specifically Germany.

III.F. Looking Back and Looking Ahead

With the benefit of hindsight, which elements of our earlier analysis have held up well and which ones less so, that is, where is there room for improvement? What lessons can we draw for future analyses of similar scenarios? For example, suppose that ten years from now another large energy supply shock looms and we would like to evaluate it using quantitative macroeconomic modeling. Or suppose China invades Taiwan and a similar debate arises about the economic costs of sanctioning China. Which parts of the analytical framework described earlier will come in handy, and where does it have gaps?

In retrospect, probably the biggest gap in our earlier analysis was the omission of demand-side effects, in particular standard Keynesian aggregate demand amplification: rising energy prices drag down consumer spending and this feeds back into production and employment, which further drags down consumption, and so on.[38] Direct empirical evidence for this type of Keynesian multiplier mechanism is hard to come by because it is

36. Bruegel, "European Natural Gas Imports," https://www.bruegel.org/dataset/european-natural-gas-imports.

37. Bruegel, "European Natural Gas Demand Tracker," https://www.bruegel.org/dataset/european-natural-gas-demand-tracker.

38. As noted in section II.F, our model did include the standard flexible-price demand-side effect that higher energy prices erode purchasing power and erode consumer welfare.

concerned with general equilibrium effects and we have not come up with a convincing empirical strategy for isolating them during this particular episode.

However, there are two reasons to believe that such effects are important in practice and should be included in full-blown analyses of negative energy supply shocks. First, this mechanism is operational in standard macroeconomic models with nominal rigidities that are consistent with empirical evidence on household consumption behavior, in particular HANK models that are consistent with the large observed marginal propensities to consume.[39]

Second, empirical analyses of past energy shocks (typically oil shocks) using time series data have documented patterns consistent with demand-side effects, in particular that these shocks primarily affected the economy through a disruption in consumer spending on goods and services other than energy (Hamilton 2008, 2009, 2013; Edelstein and Kilian 2009). For example, Hamilton (2009, 2013) shows that one of the key responses seen following the five historical oil shocks was a decline in car purchases, and argues that this accounted for a large share of the drop in GDP in the five quarters following the shocks. Hamilton (2013, 262) concludes that "combining these changes in spending with traditional Keynesian multiplier effects appears to be the most plausible explanation for why oil shocks have often been followed by economic downturns." If such demand-side amplification was important following the past oil shocks, one would expect it to also have been operational following the German economy's cutoff from Russian gas.

An interesting question is why Germany's 2022 cutoff from Russian gas appears to have been less costly than the oil shocks of the 1970s.[40] Three candidate explanations are as follows. First, both in the 1970s and today, oil plays a more important role in the global economy than natural gas, and therefore, the oil shocks were simply larger shocks. To show this, online appendix figure B.5, panel (a), compares the evolution of world oil expenditures as a share of world GDP to those on natural gas since the 1970s.

39. See Bayer, Kriwoluzky, and Seyrich (2022), Bayer and others (2023), Pieroni (2023), and Auclert and others (2023) for analyses emphasizing this mechanism.

40. It is worth noting that during the 1970s oil shocks, Germany fared better than the United States. For example, in the aftermath of the 1973–1974 oil shock, US GDP contracted by 2.5 percent (Hamilton 2009) whereas German GDP contracted by only 0.9 percent in 1975; Destatis, "Bruttoinlandsprodukt von 1950 bis 2022 im Durchschnitt 3,1 % pro Jahr gewachsen" [Gross domestic product grew an average of 3.1% per year from 1950 to 2022] https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/06/PD23_N032_81.html.

Despite larger fluctuations in both series, the oil expenditure share is consistently higher than the gas expenditure share, with oil expenditures of about 2 percent of GDP in normal times compared to 1 percent for gas. Similarly, comparing the 1970s oil and 2022 gas shocks, oil expenditure more than quadrupled from about 1.5 percent to 7 percent of world GDP in the 1970s, whereas gas expenditure rose from around 1 percent to 3.5 percent—the oil shock's peak impact was again twice as high as that of the gas shock (7 percent versus 3.5 percent).[41] Data for both Germany and the European Union as a whole paint a similar picture—see online appendix figure B.5, panel (b).[42] Tying this back to our earlier theoretical discussion, we showed in section II.A that economic costs of input supply shocks not only critically depend on the elasticity of substitution but also on the share parameter. Specifically, we showed there that (keeping $\sigma = 0.05$) an oil value for α equal to 2 percent yields output losses of 4.5 percent, which are almost twice as high as those with a gas value for α equal to 1 percent. That is, we should a priori expect the economic costs of oil shocks to be almost twice as high as those of the gas cutoff simply because the oil expenditure share is roughly twice that of gas.

Second, as noted in section II.D, structural change means that manufacturing now accounts for a smaller share (only about a quarter) of economic activity than in the past. Third, households' use of oil and gas differ in ways that could explain why high oil prices appear to be a stronger drag on consumer spending than high gas prices. Specifically, high oil prices affect consumers primarily via high petrol prices, whereas high gas prices affect heating costs. Petrol prices are much more closely tied to spot market prices than heating costs, which are determined by relatively longer-term contracts. Petrol costs are arguably also more salient and may thus affect consumer spending and confidence more strongly.[43]

41. Note that the oil shock was also much more persistent. Consistent with our numbers, Baqaee and Farhi (2019, fig. 7) calculate that the global expenditure on crude oil as a share of world GDP was around 2 percent and quadrupled to 8 percent in the 1970s.

42. Also recall online appendix figure B.3, which showed an increase in Germany's gas expenditure share in GNE from 1 percent to 4 percent. The larger impact for Germany in figure B.3 than in figure B.5, panel (b), is primarily due to the use of higher frequency monthly data in figure B.3, with monthly gas prices showing a larger peak than the yearly data in figure B.5, panel (b).

43. Finally, a potential alternative explanation is that many oil shocks appear to be strongly temporally correlated with large monetary policy shocks (Hoover and Perez 1994; Nakamura and Steinsson 2018), implying that inference about the separate effects of either type of shock is complicated.

On the flip side of paying more attention to Keynesian demand amplification, future analyses should probably spend relatively less time and effort quantifying the cascading effects discussed in section II.D. This is because the data instead showed a substantial decoupling of overall industrial production from that in a few energy-intensive sectors like chemicals and glass, the polar opposite of cascading effects. The focus on cascading effects in our original paper (Bachmann and others 2022b) was due to these effects being a central (or perhaps even *the* central) concern in the German public debate back in the spring of 2022. In retrospect, this also reflected that lobbyists are skilled at shifting public debates, in particular, taking advantage of the fact that the "Leontief logic"—everything drops proportionately—is extremely intuitive for nonspecialists. The absence of cascading effects and the strength of the observed decoupling between energy-intensive production and the rest is interesting from an economic perspective. Once more, when the granular data on industrial production and gas usage become available, it would be interesting to see how exactly this decoupling played out in practice.

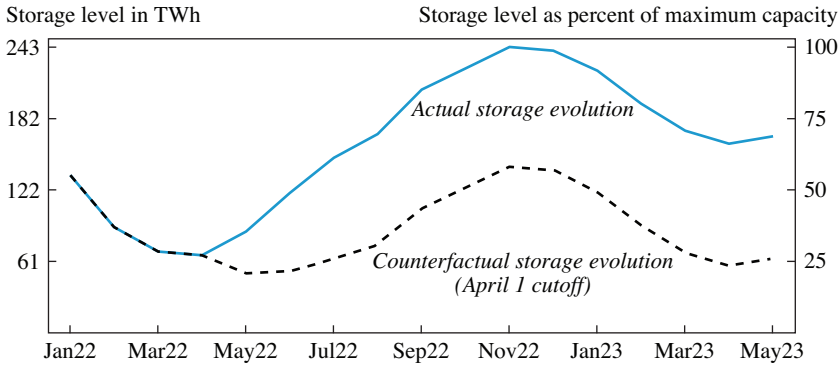
IV. Could Germany Have Withstood an Earlier Cutoff as Well?

To what extent did the timing of the cutoff matter for these benign economic outcomes? It is clear now that the cutoff from Russian gas that Germany experienced in the summer of 2022 had moderate and manageable economic consequences, and that the country even exited the winter with substantial gas reserves of around 65 percent (see figure 10 below). But it is an open question whether Germany would have made it through the winter with an earlier cutoff, possibly as early as April 2022, which would have left only a few weeks for preparations.

A prominent line of argument is that the additional months from April to August, during which Germany continued to import and stockpile Russian gas, were decisive to fill storage capacity sufficiently to get through the winter. Without those Russian imports, the argument goes, with an immediate severance from Russian energy starting in April 2022, shortages, rationing, and high economic costs would have ensued.

We here provide some simple counterfactual calculations to answer this question, taking April 1, 2022, as the hypothetical cutoff date. We ask the following simple questions: In retrospect, would Germany still have had gas left in its gas storage facilities at the end of the 2022–2023 winter if the country had stopped importing Russian gas on April 1, 2022, rather than

Figure 10. Counterfactual Storage Evolution with Gas Cutoff at the End of March 2022



Source: Bruegel; and authors' calculations.
Note: See online appendix C for details on sources and the construction of the series for counterfactual storage evolution.

continuing to import and stockpile Russian gas until the end of August 2022? Or would Germany have run out of gas in the middle of the winter?

Figure 10 presents a simple counterfactual scenario that answers this question. The solid line plots the actual observed storage evolution including Russian gas imports after March 2022. The dashed line plots the counterfactual storage evolution in the event of an April import stop calculated from combining data on Russian gas imports and the observed storage evolution (see the explanation below and in the online appendix). The key takeaway is that even with an April 1 gas cutoff, Germany would still have exited the winter with gas storages that are 25 percent full. In other words, Germany would have been able to cope with an earlier April embargo.

The following simple calculation explains this result. We compute the cumulative observed imports of Russian gas over the period from April to August 2022, taking into account imports via third countries as well as reexports (see online appendix for details) and compare this number to the amount of gas left in German storages at the end of the 2022–2023 heating period. The idea is simple: holding consumption and other gas supplies constant, if Germany exited the winter with more gas left in its storages than these cumulative imports, then Germany would not have run out of gas even with an April import stop from Russia; in contrast, if gas reserves at the end of the winter were less than these cumulative imports, Germany may have run out of gas without these imports.

Germany had imported about 100 TWh of Russian gas since April 2022, which is about 10 percent of the typical annual gas consumption in previous years or about 40 percent of maximum storage capacity.[44] On the other hand, Germany had about 160 TWh of gas left in its storage facilities, which is about 16 percent of typical annual consumption or about 65 percent of storage capacity. Therefore, even with an April 1 gas cutoff, Germany would still have emerged from the winter with gas storages that were 25 percent full (65% − 40% = 25%), which is exactly the number plotted in figure 10—see the data point for April 2023.

In fact, the 25 percent storage level implied by this simple counterfactual calculation should be viewed as a lower bound, that is, Germany would have arguably emerged from the winter with higher gas storage levels. First, our counterfactual calculation holds constant German gas consumption, that is, it assumes that even with gas supplies falling much more substantially and storage levels being considerably lower before the start of the winter, consumption would have been unchanged relative to its actual time path. This assumption is unrealistic: instead, with lower supplies and storage levels, further demand reduction would likely have occurred.[45] Second, there was a time period in October and November 2022 during which German gas storages were virtually full and therefore gas imports were constrained by a lack of storage capacity—nowhere to put this gas. In fact, gas storages not just in Germany but all over Europe were so full at this point that this resulted in large numbers of LNG tankers queuing off Europe's coasts, unable to unload.[46] While our calculation provides a lower bound on gas storage levels at the end of the 2022–2023 winter, we view it as useful because of its simplicity.

44. For Germany-wide maximum storage capacity we use 246 TWh, based on the fact that storages were completely filled by early November 2022 with 246 TWh. Similarly, there is a question as to what the minimum storage level is at which storages can still operate efficiently. The lowest historical storage filling level was only 35 TWh of working gas in March 2018, significantly below the 60 TWh in our counterfactual scenario, and even at 35 TWh storages still contained significant volumes of cushion gas that could have been extracted in an emergency situation; Gas Infrastructure Europe (GIE), "Aggregated Gas Storage Inventory (AGSI) Data Overview," https://agsi.gie.eu/data-overview/graphs/DE.

45. This mechanism, additional demand reduction, would have likely been a particularly powerful force toward higher storage levels. This is because German gas storages are small relative to typical gas demand: maximum gas storage capacity is 246 TWh, which is only about a quarter of annual gas consumption of about 1,000 TWh (Bachmann and others 2022a). Thus, even an additional demand reduction of only 2 percent would have reduced demand by 20 TWh and would have increased the storage filling level at the end of the winter from 60 TWh or 25 percent to 80 TWh or 33 percent.

46. See, for example, Rashad and Carreño (2022) and LaRocco (2022).

To construct the full time path for counterfactual storage evolution in figure 10, we further break down imports of Russian gas by month. Online appendix figure C.1 plots the results and highlights that, while Germany continued to import Russian gas through the end of August 2022, these imports were small from June onward when Russia started weaponizing gas.[47] Using these monthly data, figure 10 is then computed by subtracting the Russian imports for each month from the observed storage net inflows. Apart from our main argument that Germany would have not exhausted its gas reserves at the end of the 2022–2023 heating period, figure 10 makes another important point, namely, that gas storages are also not exhausted at any other point in time after April 2022. Put differently, the combination of gas imports from other countries and preexisting storage would have been sufficient to satisfy both industrial and household gas demand at any point in time.

In particular, contrary to the arguments of some skeptics, there was never a danger of a gas shortage immediately following an April gas cutoff. One important reason for this result is the well-known seasonality of gas demand—that gas demand is much lower in the summer. An April cutoff would have coincided with the end of the 2021–2022 heating period and thus the start of the low-demand summer period, meaning that even relatively low levels of preexisting storage would have been enough to prevent shortages and rationing. That the seasonality of gas demand means that there would be no immediate gas shortages even with a cold turkey import stop was an important argument in Bachmann and others (2022b).[48]

Although we focus on the outcomes in Germany, our counterfactual scenario considers a cutoff from Russian gas for the European Union as a whole rather than just Germany. Because the European gas market is complex and heavily interconnected, we therefore take into account not only direct imports to Germany from Russia (via the Nord Stream 1 pipeline) but also indirect imports via third countries (e.g., flows via Ukraine Transit and Czechia or Austria to Germany) as well as reexports. Thus, our series for imports from Russia includes only the gas that actually entered and was consumed or stored in Germany and would have been therefore "missing"

47. Thus, the skeptics' argument that the additional five months from April to August, during which Germany continued to import and stockpile Russian gas, were decisive for getting the country through the following winter is really an argument about two months alone, April and May.

48. Of course, an earlier import stop would likely have moved gas prices by more or earlier, or both. This would have likely resulted in higher economic costs. On the flip side, it would have also resulted in larger demand reduction as already discussed.

in the event of an earlier import stop. Our counterfactual scenario then subtracts these missing imports from total net inflows into German storages. Note that the subtracted missing imports do not include Russian gas that used to be reexported to third countries because doing so would overstate the gas shortfall by effectively assuming that, after April 1, Germany would have just reexported the same amount of gas as if nothing had happened despite being cut off from Russian gas. The online appendix contains details and discusses a number of additional considerations.

V. The Role of Luck

In any year, gas supply and gas demand are affected by numerous exogenous factors whose unpredictable realizations can noticeably ease or tighten the supply-demand balance. The most important factor is the weather (section V.A), but there are also many other important variables like accidents, strikes, and conflicts, specifically those affecting the European electricity market (section V.B), as well as the availability of LNG, which played an important role in displacing Russian gas (section V.C).

V.A. Was the 2022–2023 Winter Particularly Warm?

Heating demand and hence temperature is a main driver of gas demand in Germany. If on one cold day the average temperature falls by 1°C, the total daily gas consumption in Germany will increase by about 165 GWh. This means that, on a day with a temperature of 0°C, a 1°C change corresponds to 6–7 percent of gas consumption. Most of this temperature sensitivity of demand is due to small and household consumers.[49]

At a very basic level, the average winter temperature for Germany in the 2022–2023 winter of 2.9°C was actually slightly colder than the average temperature of 3.0°C over the four previous winters.[50] However, a more systematic analysis is required. To account for the fact that when it is already warm outside, heating demand is relatively unresponsive to temperature changes (say from 20°C to 21°C daily average), energy economists like

49. About 120 GWh higher demand per degree comes from small consumers alone in Germany on average. Numbers here are authors' own calculations based on the Eurostat data of sectoral gas demand and heating degree days.

50. Deutscher Wetterdienst, "Zeitreihen für Gebietsmittel für Bundesländer und Kombinationen von Bundesländern" [Time series for area averages for federal states and combinations of federal states], https://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/seasonal/air_temperature_mean/regional_averages_tm_winter.txt; accessed via "Mittelwerte für die einzelnen Bundesländer und für Gesamtdeutschland," https://www.dwd.de/DE/leistungen/cdc/cdc_ueberblick-klimadaten.html.

to use heating degree days (HDDs). HDDs are a measure of the severity of the cold (specifically, how much the outside temperature is below 18°C) and hence the need for heating over a specific time period. Figure 11, panel A, shows that monthly HDDs are almost perfectly correlated with monthly gas consumption.

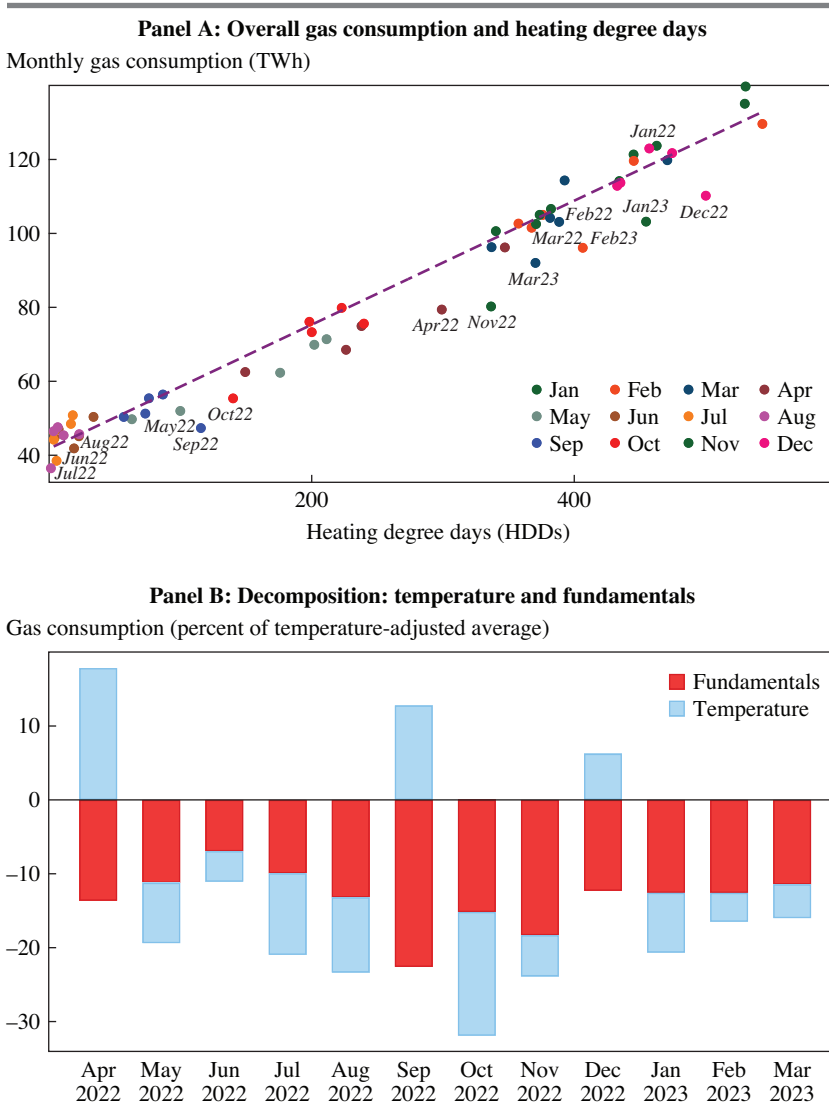
Figure 11 also shows that, following the Russian invasion of Ukraine (i.e., from March 2022), all monthly gas consumption fell below the linear trend that indicates the expected gas consumption given a month's HDDs. For example, December 2022 was particularly cold and showed a high number of 500 HDDs (in the previous five years, December had between 433 and 475 HDDs), which would normally imply 123 TWh of gas consumption. However, despite these cold temperatures, in December 2022 Germans consumed only 107 TWh.

Overall, the year 2022 had 2,736 HDDs in Germany. This can be compared to three different baselines. First, comparing it to the previous year 2021 with 3,114 HDDs makes 2022 look like a warm year. But 2021 was actually the coldest year since 2013 (as measured by HDDs), meaning that 2021 was an outlier. Second, one can compare it to the average of the previous decade of 2,933 HDDs per year. But this decadal average is not a good measure of the expected number of HDDs for 2022 either. The reason is climate change. Our third and preferred comparison accounts for this trend: using data since 1979, online appendix figure D.1 shows that the number of HDDs declined by about 14 HDDs every year. Along this long-term trend line, the expected number of HDDs in 2022 was about 2,850. Thus, with 2,736 HDDs, the year 2022 had only 114 fewer HDDs (the year was slightly less cold as measured by HDDs). Converting these 114 HDDs into gas consumption using the correlation in figure 11, panel A, implies a reduction in gas consumption of only 18 TWh or 1.8 percent of average consumption. Hence, as measured by HDDs and the implied gas demand, Germany was not particularly lucky.[51]

Taking this logic one step further, we can also decompose the observed reduction in gas consumption into a part due to temperature and another part due to "fundamentals" (i.e., factors other than temperature). For example, a baseline year with 2,850 HDDs would have implied a gas demand of 996 TWh. Compared to that, Germany's 2022 consumption of 854 TWh implied a demand reduction of 142 TWh. Hence, the 18 TWh savings from slightly milder temperatures accounted for less than 13 percent of

51. On the flip side, it is true that Germany was also not particularly unlucky. For example, a very cold winter like 2021 would have increased gas consumption by about 61 TWh.

Figure 11. Temperature-Adjusted Gas Consumption



Source: Bundesnetzagentur; and Eurostat.

Note: Gas consumption data are from Bundesnetzagentur, "Gasverbrauch Haushalts- und Gewerbekunden, wöchentlicher Mittelwert" [Gas consumption households and businesses weekly]. Data on heating degree days (HDDs) are from Eurostat (database code nrg_chdd_m). HDDs are a measure of the severity of the cold, specifically, how much the outside temperature is below 18°C, and hence the need for heating. In panel A, the line is fitted using data up to March 2022. In panel B, the reduction in gas consumption compared to the pre-2022 average is decomposed into two parts. The term "fundamental" represents the difference between actual gas consumption and its predicted value from the fitted line, while the remainder is called "temperature."

the savings, that is, the remaining 87 percent were due to fundamentals. Figure 11, panel B, uses the correlation in panel A to conduct a similar exercise for each month in the period from April 2022 to March 2023. The results show that in all but one month, mild temperatures played a minor role in accounting for reduced gas consumption (the exception is October 2022). In fact, both September and December 2022 were unusually cold but nevertheless saw substantial gas savings. These calculations confirm the results by Ruhnau and others (2023) and Roth and Schmidt (2023), who find that substantial savings happened even after controlling for temperature.

Finally, the warmer temperatures in October and November 2022 contributed disproportionately little to getting Germany through the winter. This is because the warmer temperatures (smaller number of HDDs) occurred at a time when gas storages were virtually full. Hence, higher temperatures in October and November resulted in lower gas prices but not a better preparation for the coming winter.

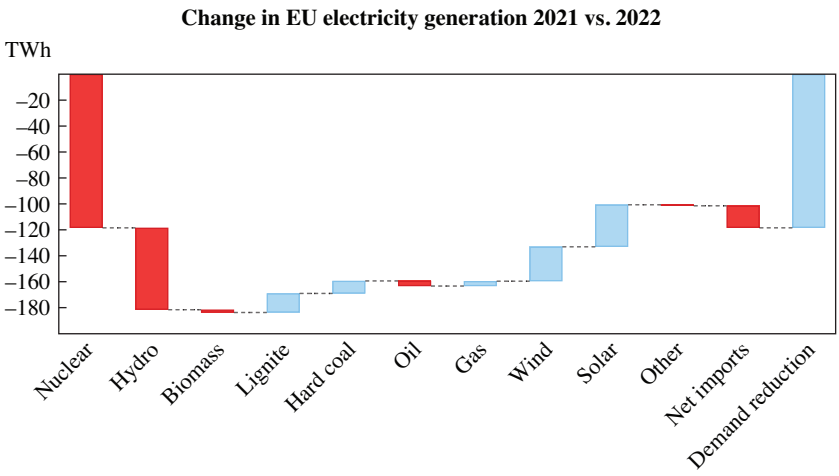
V.B. Shortfalls in Electricity Generation Prevented Fuel Switching

Different energy commodities show strong interactions. This is particularly true for natural gas and electricity. The two are direct substitutes for producing heat and a significant share of electricity is produced from natural gas. Their demand has many common drivers like weather and economic activity. Moreover gas and electricity demand and prices interact indirectly through other commodity markets, especially those for emission allowances and coal. Most importantly, even though gas-fired power plants are a relatively expensive and inefficient way of producing electricity, there are many hours each day during which electricity production relies on natural gas simply because cheaper options alone are insufficient to meet demand. Notably, because one needs about two MWh of gas to produce one MWh of electricity, the marginal cost and hence the hourly wholesale electricity price per MWh in these hours is about twice the gas price per MWh. Accordingly, developments in the gas market spill over into the wholesale electricity market, which has roughly the same annual turnover.

This high degree of interaction has two relevant implications: First, gas savings may be achieved via fuel switching in electricity production (e.g., from gas to oil or coal) or via reduced electricity consumption. Second, high gas prices have a very strong impact on electricity prices.

In 2022, however, special conditions in electricity markets meant that the first effect did not actually contribute to mitigating the gas crisis. Maintenance issues at French reactors meant that French nuclear generation in

Figure 12. Reduced Ability of the Electricity System to Alleviate the Gas Scarcity



Source: Energy-Charts.

Note: Data are based on European Network of Transmission System Operators for Electricity (ENTSO-E).

2022 was 82 TWh (or 22 percent) below the already low 2021 values. Moreover, the long-planned shutdown of three German reactors at the end of 2021 reduced power generation by 32 TWh and a drought reduced hydro generation in the European Union by 82 TWh compared to 2021. Reduced nuclear and hydro generation in 2022 meant that the European Union lacked about 180 TWh (7 percent) of its low-cost electricity supplies (see figure 12). Replacing this electricity production shortfall with gas-fired generation—which is often the marginal fuel in the northwest European power market—would have required burning about 360 TWh more natural gas in power plants.[52] As a result, the European electricity system, which would normally have served as a substantial buffer to gas supply issues by switching to using more coal and reducing electricity demand, was already extremely stretched due to its own internal problems. Therefore, despite the largest gas crisis in recent history, Europe actually increased gas consumption in the power sector slightly from 432 TWh to 436 TWh instead of decreasing it as predicted by economic theory.[53] These elements of

52. As gas-fired power plants have an efficiency of about 50 percent in transforming the heating energy of natural gas into electric energy, it takes about 400 TWh of gas to produce about 200 TWh of electricity.

53. The data on electricity generation used in this paragraph are from Energy-Charts.

bad luck also explain the very small contribution of power generation to demand reduction in Germany in table 2.

V.C. The Role of LNG

Whether the situation in global LNG markets was favorable to weathering the gas cutoff is a difficult question. It is clear that massive EU LNG imports induced higher global LNG prices and hence triggered supply extension and demand reduction in other markets. But whether lower Asian gas demand in 2022 was driven primarily by unexpected local factors (e.g., the slower than expected post-COVID-19 recovery) or whether this low demand was a reaction to the very high LNG prices is hard to disentangle empirically.[54]

Moreover, in June 2022, the Freeport LNG plant in the United States, the fourth-largest LNG liquefaction plant in the world, was put out of action by a fire and only restarted loading cargoes in mid-February 2023. Had it not been dysfunctional, this plant would have been able to liquify more than 100 TWh of US natural gas.[55]

In conclusion, the bad luck elements actually exceeded the good luck ones over the last year. The role of good luck in getting Germany through the winter has been considerably overstated in the popular debate.

VI. Political Economy of Decision Making in Times of Crisis

Some of the most important lessons from the great German gas debate concern the political economy of decision making in times of crisis. While some of these lessons are linked to specific features of the German corporatist model of close coordination between government, business associations, and trade unions, others likely extend beyond the narrow German context and are important to be reflected upon. In particular, the tensions between China and Taiwan could well lead to comparable developments where policymakers might have to navigate similar trade-offs between business interests and foreign policy objectives. In the German case, the most

54. Asian LNG imports decreased from 273 MT LNG to 252 MT LNG, whereby China alone reduced by 16 MT according to GIIGNL (2023).

55. Freeport has a liquefaction capacity of about 20 billion cubic meters per year, hence more than 100 TWh in the eight months of its dysfunctionality. Enerdata, "JERA Will Buy 25.7% of the Freeport LNG Project (US) for US$2.5bn," November 17, 2021, https://www.enerdata.net/publications/daily-energy-news/jera-will-buy-257-freeport-lng-project-us-us25bn.html.

important insights have to do with the outsized role of business leaders and their associations in times of acute crisis. One does not have to agree with Adam Smith's (1776, 16) famous quip that congregations of businessmen often end in a "conspiracy against the public" to conclude from the recent experience that geopolitical dynamics can bring specific incentive problems for profit-maximizing business leaders.

When the discussion about Germany's vulnerabilities began after the Russian invasion, policymakers did not turn to academics but to business leaders and their associations for advice. The key interlocutors were representatives of the most affected industries such as the energy and chemicals sectors, refineries, and other industrial companies. This was primarily due to policymakers' concern to understand the practical implications of a cutoff from Russian gas and what this would mean for operations "on the ground."

While understandable, this also meant that the very industries that had made large commercial bets on Russian gas became the main interlocutors, thereby blurring commercial interests and political influence once again. Business leaders had a clear incentive to talk up the dependence on Russian gas in their interaction with policymakers in Berlin, thereby making a stronger political and military reaction by the German government less likely and indirectly increasing the chances of continued access to cheap Russian gas for their companies. Most CEOs and leaders of industry associations were outspoken that the consequences of a cutoff from Russian gas would be catastrophic. The feedback was that the dependence was extremely high, and that in the short run, no alternatives existed so that production cuts coupled with cascading effects down the production chain would be inevitable consequences of a gas cutoff. Union representatives, mainly concerned with potential job losses, were quick to support the position of business leaders.

The CEO of the German chemicals giant BASF, Martin Brudermüller, became a particularly vocal advocate of the dependency camp, predicting that a cutoff from Russian gas "could bring the German economy into its worst crisis since the end of World War II and destroy our prosperity" and asking, "Do we knowingly want to destroy our entire economy?" (Brankovic and Theurer 2022, par. 4 and 12).

Yet in some cases, the very same businesses whose CEOs had denied any short-run possibility of gas savings or substitution announced substantial reductions in gas usage only a few weeks later or found substitution possibilities of the very kind that had been discussed in the public debate.

For instance, having warned of a shutdown of its huge plant in Ludwigshafen, BASF announced soon thereafter that its Verbund system would be able to run with half the usual gas supplies and that gas-intensive ammonia production could be transferred to a BASF plant in the United States and imported from there.[56]

To what extent these early statements shaped Germany's initial hesitancy to supply Ukraine with more advanced weapons quickly is a question that future historians will have to address. But it is worth highlighting that neither economic arguments on demand responses to price increases and substitution possibilities, nor empirical studies from previous interruptions of energy supplies in other countries, carried enough weight to be a counterweight to the presumed real-world knowledge of business leaders, as conflicted as they might have been. Both theoretical and empirical reasoning of economists was deemed much less relevant than the judgment of company CEOs, a major reason likely being the potential political costs of going against the explicit advice of company and union leaders.[57]

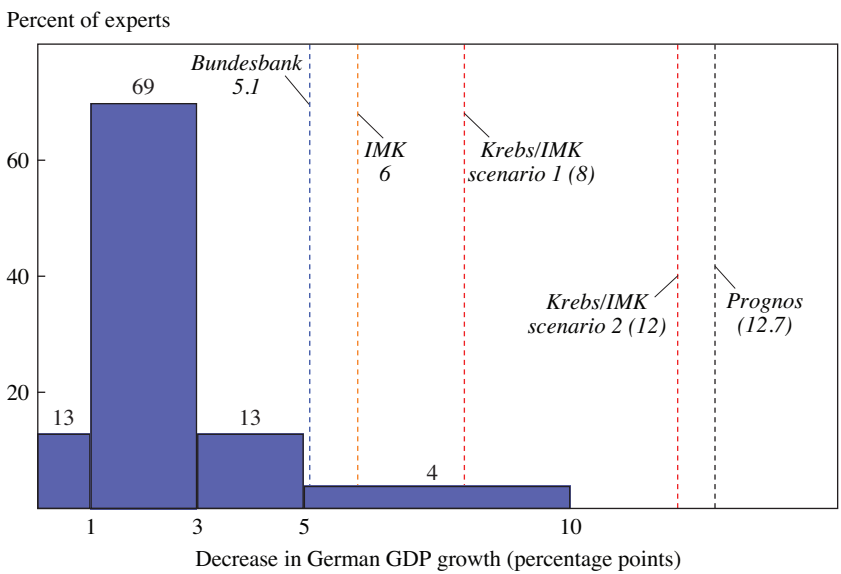
A second important lesson relates to the strategic use of think tanks associated with business and union interests to increase the uncertainty of cost estimates.[58] In practice, individual industry and union lobbies would pay for additional studies that arrived at high-cost estimates using extreme assumptions. Figure 13 contrasts the prediction of some of these studies to an April 2022 survey of academic economists about the likely effects of a Russian gas cutoff. Although the bulk of responses of academic economists were clustered in a reasonably narrow range up to 5 percent of GDP,

56. While BASF had been publicly stating that half of its normal gas supplies would be sufficient as early as March 2022, one particularly clear version is an investor conference call presentation from July 2022 stating, "Continued operation at Ludwigshafen site is ensured down to 50 percent of BASF's maximum natural gas demand" (BASF 2022; and case 18 in online appendix E). For ammonia substitution via imports, see section III.B and cases 2 and 15 in online appendix E.

57. After criticizing the "irresponsible use of mathematical models" on the *Anne Will* TV show (see introduction), Chancellor Scholz added, "I don't know absolutely anyone in business who doesn't know for sure that [entire branches of industry shutting down in the event of a gas cutoff] would be the consequences"; see the transcript and English translation available at https://benjaminmoll.com/Scholz/.

58. Banerjee and Duflo (2019) warn against the role of economists representing special interests in the public debate. Two special interest–financed think tanks stand out in Germany: the Institut der deutschen Wirtschaft (IW), which is financed by various industrial lobbies, and the Institut für Makroökonomie und Konjunkturforschung (IMK), which is largely financed by the German trade union federation DGB.

Figure 13. Studies Financed by Special Interest Groups Predicted Much Larger GDP Losses than Academic Economists



Source: Centre for Macroeconomics (CFM).

Note: The histogram represents the answers by European academic economists to question 2 in the April 2022 CFM survey on the effects of an embargo on Russian gas, "By how much would an immediate EU-wide import ban on Russian gas reduce German GDP growth per annum in 2022-3, in percentage points (pp), if the government offset the costs with a well targeted fiscal policy?" The dashed lines plot the estimates by Deutsche Bundesbank (2022), Behringer and others (2022), Krebs (2022), and Prognos (2022). For context, IMK is a union-financed think tank, the Krebs study was paid for by the German trade union federation DGB, and the Prognos study was paid for by a business association.

the studies financed by special interest groups produced much larger numbers of up to 12.7 percent of lost output.[59]

While the economic debate focused on the content of these studies and the underlying extreme assumptions, their political goal was a different one. By substantially broadening the range of potential cost estimates of a cutoff from Russian gas, they undermined public confidence in the reliability of

59. Centre for Macroeconomics (CFM), "Effects of an Embargo on Russian Gas," The CFM Surveys, https://www.cfmsurvey.org/copy-of-survey-2022-05. For reference, figure 13 also plots the largest cost estimate not financed by a special interest group, a 5.1 percent GDP drop predicted by Deutsche Bundesbank (2022). It is worth pointing out that Bundesbank cost estimates significantly exceeded those of other comparable institutions. For example, three IMF studies—Lan, Sher, and Zhou (2022), Albrizio and others (2022), and Di Bella and others (2022)—predicted more moderate economic losses of up to 3 percent of GDP. Also see the follow-up study by Albrizio and others (2023).

any cost estimate and increased uncertainty about the consequences in the eyes of the public. The impression remained that even experts could not agree about this matter so the prudent thing was to conclude that we simply cannot know how bad things can possibly get—reinforcing the approach taken by policymakers. Given that the uncertainty about economic estimates was so large, they could be dismissed altogether and other sources of information—such as contacts with company leaders—could be considered reliable.

Ultimately, the main effect of these academically questionable studies that arrived at extremely high economic costs was to create the impression of uncertainty, allowing policymakers to dismiss academic advice as too uncertain. A good example of this is captured in the following quote by Jörg Kukies, the head of the Economics Division in the Chancellor's Office in Berlin: "We will never ever be able to determine whether this has a 2 percent or 10 percent GDP impact. . . . We are simply trying to take the pragmatic middle course because we do not know and cannot know [what the effect would be of] such an abrupt termination" (Kukies 2022, minute 8:55 and 10:13).[60]

VII. Conclusion

It was primarily the economy's ability to adapt in combination with the insurance offered by trade and (some) good economic policymaking that blunted Putin's energy weapon: as prices rose, German producers and households reduced demand and substituted away from natural gas, the country quickly sourced alternative gas supplies, and policymakers implemented well-designed policies to support households and firms that maintained price signals to encourage gas to go to the sectors and countries where it was most needed.

The cutoff from Russian gas is an unusually clear case of how consumers and producers react when an important input (here natural gas) becomes scarce and expensive. As new data covering the 2022–2023 time period are starting to become available, future work should examine in more detail how this significant shock propagated across sectors, regions, and countries as well as its distributional effects. This work could also use the gas cutoff

60. The original German is "Wir werden es nie und nimmer entscheiden können, ob das jetzt 2% oder 10% BIP-Einfluss hat. . . . Wir versuchen einfach den pragmatischen Mittelweg zu gehen, weil wir nicht wissen und nicht wissen können [was der Effekt ist] bei einem so abrupten Abbruch."

as a natural experiment to identify and estimate various elasticities that will be relevant in other contexts. Prime examples are questions regarding the green transition, in particular projecting the economic impact of rising carbon prices, which will affect similar sectors of the economy as the gas shock. There are, however, limits to the comparison. For example, decarbonization will imply a continuous and universal decrease in the supply of emission permits, while the gas crisis cut out only one major gas supplier (Russia).

The main rationale for sanctioning Russian energy exports has always been simple, namely, that these exports represent an important source of fiscal revenues for the Russian state—money that is then used to wage war in Ukraine. As Oleg Itskhoki has put it: "Each marginal euro received [by Russia] from energy exports to Europe contributes exactly one euro to the war, as simple as that."[61]

Despite this clear rationale for sanctioning Russian energy exports, Western countries opted for a cautious approach and such sanctions did not begin in earnest until the EU crude oil embargo took effect in December 2022, almost ten months after the start of the war. Sanctions on gas exports have still, to this day, been absent from any sanctions packages. This delayed and cautious implementation of energy sanctions contributed to Russia earning record export revenues in 2022 and likely to its ability to wage war in Ukraine. For example, Babina and others (2023) argue that even though the EU oil embargo only came into effect in December 2022, it has already materially affected Russian export revenues and, furthermore, that an earlier introduction of the EU oil embargo and the G7 price cap in the immediate aftermath of the invasion could have reduced Russia's oil export earnings by up to $50 billion or about one-third.

Naturally, just like Germany substituted and adapted in the face of the gas cutoff, Russia has also been substituting and adapting in the face of Western sanctions. The power of substitution cuts both ways. However, the Russian government's strong reliance on fiscal revenues from energy

61. X (Twitter), April 8, 2022, https://twitter.com/itskhoki/status/1512508687641763844?s=20. A particularly good exposition of the case for energy sanctions is by Guriev and Itskhoki (2022). Opponents of the energy embargo idea have often argued that Russian war expenditures would be unaffected because the Russian government can print its own money and therefore does not need to rely on export revenues. A good rebuttal of this argument is made by Hanno Lustig: "Suppose we did a helicopter drop of dollars in Red Square in Moscow. If no one bothers to pick them up, then export curbs are indeed irrelevant. Not a likely outcome of this experiment"; X (Twitter), June 4, 2022, https://twitter.com/HannoLustig/status/1533000546659012608?s=20.

exports does mean that the situation is asymmetric and that export sanctions likely bite.[62]

One manifestation of declining export revenues due to energy sanctions has been the ruble's depreciation throughout the spring and summer of 2023 (Itskhoki and Mukhin 2022; Lorenzoni and Werning 2023). This has already forced hard choices on Russian policymakers with the central bank recently implementing significant interest rate hikes (Guriev 2023).

Keeping Russia's natural gas exports out of the sanctions regime generates substantial revenues for the Russian state—some €200 million per week (Levi 2023). Not sanctioning the financial institutions used for the corresponding payments, specifically Gazprombank, is similarly problematic. Apart from the unsanctioned gas exports contributing to Russia's war effort, Europe effectively allowed Russia to decide on the price and volume of these exports to individual destination countries, thereby creating divisions between countries that still receive Russian gas via pipeline (e.g., Austria and Hungary) or LNG (e.g., Spain) and those that do not. As Europe will continue to use natural gas for at least two decades and Russia's gas export infrastructure to Europe is still very potent, Europe should consider taking advantage of the historically low flows to establish joint political control over gas flows from Russia rather than buying cheaply produced gas at high prices.

The failure by Western countries to implement sanctions sooner and more decisively represents a major missed opportunity to stand up to Putin and help avert enormous human suffering in Ukraine. There are good arguments that the West should tighten its sanctions regime against Russia, including on natural gas and oil, and avoid making the same mistakes in future similar crises.

ACKNOWLEDGMENTS We are heavily indebted to Ben McWilliams at Bruegel for conducting much of the data work. We are also grateful to Jim Hamilton, Tarek Hassan, Dmitry Mukhin, and Jón Steinsson for useful comments and to Sven Eis, Marina Feliciano, and Seyed Hosseini-Maasoum for excellent research assistance.

62. In the words of former US senator John McCain: "Russia is a gas station masquerading as a country. It's kleptocracy. It's corruption. It's a nation that's really only dependent upon oil and gas for their economy, and so economic sanctions are important." Transcript of McCain's interview on CNN's *State of the Union with Candy Crowley* is available at https://cnnpressroom.blogs.cnn.com/2014/03/16/sen-john-mccain-u-s-needs-fundamental-reassessment-of-russia-relationship/.

References

Albrizio, Silvia, John C. Bluedorn, Christoffer Koch, Andrea Pescatori, and Martin Stuermer. 2022. "Market Size and Supply Disruptions: Sharing the Pain of a Potential Russian Gas Shut-off to the European Union." Working Paper 2022/143. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/07/18/Market-Size-and-Supply-Disruptions-Sharing-the-Pain-of-a-Potential-Russian-Gas-Shut-off-to-520928.

Albrizio, Silvia, John C. Bluedorn, Christoffer Koch, Andrea Pescatori, and Martin Stuermer. 2023. "Sectoral Shocks and the Role of Market Integration: The Case of Natural Gas." *American Economic Association Papers and Proceedings* 113: 43–46.

Atkeson, Andrew, and Patrick J. Kehoe. 1999. "Models of Energy Use: Putty-Putty versus Putty-Clay." *American Economic Review* 89, no. 4: 1028–43.

Auclert, Adrien, Hugo Monnery, Matthew Rognlie, and Ludwig Straub. 2023. "Managing an Energy Shock: Fiscal and Monetary Policy." Working Paper 31543. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w31543.

Babina, Tania, Benjamin Hilgenstock, Oleg Itskhoki, Maxim Mironov, and Elina Ribakova. 2023. "Assessing the Impact of International Sanctions on Russian Oil Exports." Working Paper. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4366337.

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Ben McWilliams, and others. 2022a. "How It Can Be Done." Policy Brief 34. Bonn: ECONtribute.

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick. 2022b. "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia." Policy Brief 28. Bonn: ECONtribute.

Banerjee, Abhijit V., and Esther Duflo. 2019. *Good Economics for Hard Times*. New York: PublicAffairs.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87, no. 4: 1155–203.

Baqaee, David Rezza, and Emmanuel Farhi. 2024. "Networks, Barriers, and Trade." *Econometrica* 92, no. 2: 505–41.

BASF. 2022. "Analyst Conference Call Q2 2022." July 27. Slides. https://www.basf.com/global/documents/en/investor-relations/calendar-and-publications/presentations/2022/BASF_Charts_Analyst_Conference_Call_Q2-2022.pdf.assetdownload.pdf.

Bayer, Christian, Alexander Kriwoluzky, and Fabian Seyrich. 2022. "Stopp russischer Energieeinfuhren würde deutsche Wirtschaft spürbar treffen, Fiskalpolitik wäre in der Verantwortung" [Stopping Russian energy imports would have a noticeable impact on the German economy, and fiscal policy would be responsible]. Technical Report 80. Berlin: DIW Aktuell.

Bayer, Christian, Alexander Kriwoluzky, Fabian Seyrich, and Antonia Vogel. 2023. *Makroökonomische Effekte der finanz- und wirtschaftspolitischen Maßnahmen der Entlastungspakete I–III sowie des wirtschaftlichen Abwehrschirms* [Macroeconomic effects of the fiscal and economic policy measures of the relief packages I–III and the economic defense shield]. Berlin: DIW Berlin.

Behringer, Jan, Sebastian Dullien, Alexander Herzog-Stein, Peter Hohlfeld, Katja Rietzler, Sabine Stephan, Thomas Theobald, Silke Tober, and Sebastian Watzka. 2022. *Ukraine-Krieg erschwert Erholung nach Pandemie* [Ukraine war makes recovery after pandemic difficult]. Technical Report 174. Düsseldorf: Institut für Makroökonomie und Konjunkturforschung.

Berger, Eva, Sylwia Bialek, Niklas Garnadt, Veronika Grimm, Lars Other, Leonard Salzmann, Monika Schnitzer, Achim Truger, and Volker Wieland. 2022. "A Potential Sudden Stop of Energy Imports from Russia: Effects on Energy Security and Economic Output in Germany and the EU." Working Paper 01/2022. Wiesbaden: German Council of Economic Experts. https://www.sachverstaendigenrat-wirtschaft.de/fileadmin/dateiablage/Arbeitspapiere/Arbeitspapier_01_2022.pdf.

Bingener, Reinhard, and Markus Wehner. 2023. *Die Moskau Connection: Das Schröder-Netzwerk und Deutschlands Weg in die Abhängigkeit* [The Moscow connection: The Schröder network and Germany's path to dependency]. Munich: C. H. Beck.

Brankovic, Maja, and Marcus Theurer. 2022. "BASF Chef im Interview: 'Wollen wir sehenden Auges unsere gesamte Volkswirtschaft zerstören?'" [BASF boss in an interview: "Do we want to destroy our entire national economy?"] *Frankfurter Allgemeine Zeitung*. March 31. https://www.faz.net/aktuell/wirtschaft/unternehmen/basf-chef-warnt-vor-gas-embargo-schaeden-fuer-deutsche-volkswirtschaft-17925528.html.

Bundesregierung. 2022a. "Reduction of Value Added Tax on Gas." September 14. https://www.bundesregierung.de/breg-en/news/tax-reduction-gas-2126308.

Bundesregierung. 2022b. "Energy Price Brakes Are Entering into Effect." December 24. https://www.bundesregierung.de/breg-en/news/energy-price-brakes-2156430.

Bundesregierung. 2023. *Bericht der Bundesregierung zur Wirkung der Preisbremsen* [Federal government report on the effect of price controls]. Berlin: Author. https://www.bmwk.de/Redaktion/DE/Downloads/B/20230816-bericht-wirkung-preisbremsen.pdf?__blob=publicationFile&v=8.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi. 2021. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." *Quarterly Journal of Economics* 136, no. 2: 1255–321.

Carvalho, Vasco M., and Alireza Tahbaz-Salehi. 2019. "Production Networks: A Primer." *Annual Review of Economics* 11: 635–63.

Caselli, Francesco, Miklós Koren, Milan Lisicky, and Silvana Tenreyro. 2019. "Diversification through Trade." *Quarterly Journal of Economics* 135, no. 1: 449–502.

Destatis. 2023. "Bruttoinlandsprodukt von 1950 bis 2022 im Durchschnitt 3,1 % pro Jahr gewachsen" [Gross domestic product grew an average of 3.1% per year

from 1950 to 2022]. Press release N 032. June 1. https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/06/PD23_N032_81.html.

Deutsche Bundesbank. 2022. *Zu den möglichen gesamtwirtschaftlichen Folgen des Ukrainekriegs: Simulationsrechnungen zu einem verschärften Risikoszenario* [On the possible macroeconomic consequences of the Ukraine war: simulation calculations for an intensified risk scenario]. Monthly Report, April. Frankfurt: Author.

Deutscher Bundestag. 2022. "Antwort der Bundesregierung auf die Kleine Anfrage der Fraktion der CDU/CSU" [The federal government's response to the small question from the CDU/CSU parliamentary group]. Drucksache 20/2444.

Di Bella, Gabriel, Mark J. Flanagan, Karim Foda, Svitlana Maslova, Alex Pienkowski, Martin Stuermer, and Frederik G. Toscani. 2022. "Natural Gas in Europe: The Potential Impact of Disruptions to Supply." Working Paper 2022/145. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/07/18/Natural-Gas-in-Europe-The-Potential-Impact-of-Disruptions-to-Supply-520934.

Dullien, Sebastian, and Isabella M. Weber. 2022. "Mit einem Gaspreisdeckel die Inflation bremsen" [Putting the brakes on inflation with a gas price cap]. *Wirtschaftsdienst* 3: 154–55.

Edelstein, Paul, and Lutz Kilian. 2009. "How Sensitive Are Consumer Expenditures to Retail Energy Prices?" *Journal of Monetary Economics* 56, no. 6: 766–79.

European Commission. 2022. "Save Gas for a Safe Winter: Commission Proposes Gas Demand Reduction Plan to Prepare EU for Supply Cuts." Press release, July 20. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4608.

Eurostat. 2023. "GDP up by 0.3 Percent and Employment up by 0.2 Percent in the Euro Area." Euroindicators, August 16. https://ec.europa.eu/eurostat/web/products-euro-indicators/w/2-16082023-ap.

ExpertInnen-Kommission Gas Wärme. 2022. *Sicher durch den Winter: Abschlussbericht* [Safe through the Winter: Final Report]. Berlin: Bundesministerium für Wirtschaft und Klimaschutz.

Gil Tertre, Miguel. 2023. "Structural Changes in Energy Markets and Price Implications: Effects of the Recent Energy Crisis and Perspectives of the Green Transition." ECB Forum on Central Banking. Frankfurt: European Central Bank. https://www.ecb.europa.eu/pub/conferences/ecbforum/shared/pdf/2023/Gil_Tertre_paper.pdf.

Groupe International des Importateurs de Gaz Naturel Liquéfié (GIIGNL). 2023. *The LNG Industry GIIGNL Annual Report*. Neuilly-sur-Seine: Author.

Guriev, Sergei. 2023. "The Russia Sanctions Are Working." Project Syndicate, August 22. https://www.project-syndicate.org/commentary/ruble-decline-growing-budget-deficit-funding-ukraine-war-by-sergei-guriev-2023-08.

Guriev, Sergei, and Oleg Itskhoki. 2022. "The Economic Rationale for Oil and Gas Embargo on Putin's Regime." Working Paper. https://sanctions.kse.ua/wp-content/uploads/2022/09/The-Economic-Rationale-for-Oil-and-Gas-Embargo-on-Putins-Regime.pdf.

Hamilton, James D. 2008. "Oil and the Macroeconomy." In *The New Palgrave Dictionary of Economics*, 2nd edition, edited by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan.

Hamilton, James D. 2009. "Causes and Consequences of the Oil Shock of 2007–08." *Brookings Papers on Economic Activity*, Spring: 215–61.

Hamilton, James D. 2013. "History of Oil Shocks." In *Routledge Handbook of Major Events in Economic History*, edited by Randall E. Parker and Robert Whaples. London: Routledge.

Höltschi, René. 2022. "Erdgas ist sündteuer, doch die Geschäfte blühen (noch): Die Gaskrise am Beispiel von BASF" [Natural gas is extremely expensive, but business is (still) flourishing: The gas crisis using the example of BASF]. *Neue Zürcher Zeitung*, July 27. https://www.nzz.ch/wirtschaft/chemiekonzern-basf-das-gas-wird-teuer-doch-die-geschaefte-bluehen-noch-ld.1695326.

Hoover, Kevin D., and Stephen J. Perez. 1994. "Post Hoc Ergo Propter Once More an Evaluation of 'Does Monetary Policy Matter?' in the Spirit of James Tobin." *Journal of Monetary Economics* 14, no. 1: 47–74.

Houthakker, Hendrik S. 1955. "The Pareto Distribution and the Cobb-Douglas Production Function in Activity Analysis." *Review of Economic Studies* 23, no. 1: 27–31.

Institut der deutschen Wirtschaft (IW). 2022. "Gasembargo: 'Das bedeutet zweieinhalb Jahre Stillstand'" [Gas embargo: "That means two and a half years of standstill"]. Interview with Michael Hüther and Jan Schnellenbach, April 9. Cologne: Author. https://www.iwkoeln.de/presse/interviews/michael-huether-das-bedeutet-zweieinhalb-jahre-stillstand.html.

Itskhoki, Oleg, and Dmitry Mukhin. 2022. "Sanctions and the Exchange Rate." Working Paper 30009. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w30009.

Jones, Charles I. 2005. "The Shape of Production Functions and the Direction of Technical Change." *Quarterly Journal of Economics* 120, no. 2: 517–49.

Krebs, Tom. 2022. "Economic Consequences of a Sudden Stop of Energy Imports: The Case of Natural Gas in Germany." Discussion Paper 22-021. Mannheim: ZEW—Leibniz Centre for European Economic Research.

Kukies, Jörg. 2022. "Die Europäische Integration nach Corona und der Ukraine-Invasion" [The European integration after COVID-19 and invasion of Ukraine]. Speech at IMK Forum, May 4; video available at https://www.youtube.com/watch?v=A14VolEbUOU.

Lan, Ting, Galen Sher, and Jing Zhou. 2022. "The Economic Impacts on Germany of a Potential Russian Gas Shutoff." Working Paper 2022/144. Washington: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2022/07/18/The-Economic-Impacts-on-Germany-of-a-Potential-Russian-Gas-Shutoff-520931.

LaRocco, Lori Ann. 2022. "Wave of LNG Tankers Is Overwhelming Europe in Energy Crisis and Hitting Natural Gas Prices." CNBC, October 24. https://www.cnbc.com/2022/10/24/wave-of-lng-tankers-overwhelms-europe-and-hits-natural-gas-prices.html.

Levi, Isaac. 2023. "Weekly Snapshot—Russian Fossil Fuels 3 to 9 July 2023." Centre for Research on Energy and Clean Air, July 14. https://energyandcleanair.org/weekly-snapshot-russian-fossil-fuels-3-july-to-9-july-2023/.

Lorenzoni, Guido, and Iván Werning. 2023. "A Minimalist Model for the Ruble during the Russian Invasion of Ukraine." *American Economic Review: Insights* 5, no. 3: 347–56.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

McWilliams, Ben, Giovanni Sgaravatti, Simone Tagliapietra, and Georg Zachmann. 2022. "A Grand Bargain to Steer through the European Union's Energy Crisis." Policy Brief. Brussels: Bruegel AISBL.

Mendoza, Enrique G. 1995. "The Terms of Trade, the Real Exchange Rate, and Economic Fluctuations." *International Economic Review* 36, no. 1: 101–37.

Mertens, Matthias, and Steffen Müller. 2022. "Wirtschaftliche Folgen des Gaspreisanstiegs für die deutsche Industrie" [Economic consequences of the gas price increase for German industry]. *IWH Policy Notes* 2/2022. Halle (Saale): Halle Institute for Economic Research (IWH) - Member of the Leibniz Association.

Milgrom, Paul, and John Roberts. 1996. "The LeChatelier Principle." *American Economic Review* 86, no. 1: 173–79.

Moll, Benjamin. 2022. "C. Review of Other Studies: No Single Study with Deviation of Yearly GDP from Baseline Larger than 5.3%, No Recession with GDP Drop Larger than 2.5%." Supplement to Bachmann and others (2022b). https://benjaminmoll.com/RussianGas_Literature.

Nakamura, Emi, and Jón Steinsson. 2018. "Identification in Macroeconomics." *Journal of Economic Perspectives* 32, no. 3: 59–86.

Oberfield, Ezra, and Devesh Raval. 2021. "Micro Data and Macro Technology." *Econometrica* 89, no. 2: 703–32.

Obstfeld, Maurice, and Kenneth Rogoff. 1995. "The Intertemporal Approach to the Current Account." In *Handbook of International Economics, Volume 3*, edited by Gene M. Grossman and Kenneth Rogoff. Amsterdam: North-Holland.

Pieroni, Valerio. 2023. "Energy Shortages and Aggregate Demand: Output Loss and Unequal Burden from HANK." *European Economic Review* 154: 104428.

Prognos. 2022. "Lieferausfall russischen Gases-Folgen für die deutsche Industrie" [Failure to supply Russian gas—consequences for German industry]. https://www.prognos.com/de/projekt/lieferausfall-russischen-gases-folgen-fuer-die-deutsche-industrie.

Rashad, Marwa, and Belén Carreño. 2022. "Dozens of LNG-Laden Ships Queue off Europe's Coasts Unable to Unload." Reuters, October 18. https://www.reuters.com/business/energy/dozens-lng-laden-ships-queue-off-europes-coasts-unable-unload-2022-10-17/.

Roth, Alexander, and Felix Schmidt. 2023. "Not Only a Mild Winter: German Consumers Change Their Behavior to Save Natural Gas." *Joule* 7, no. 6: 1081–86.

Ruhnau, Oliver, Clemens Stiewe, Jarusch Muessel, and Lion Hirth. 2023. "Natural Gas Savings in Germany during the 2022 Energy Crisis." *Nature Energy* 8: 621–28.

Samuelson, Paul A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Sinn, Hans-Werner. 2022. "Is Germany Sick Again?" Project Syndicate, November 25. https://www.project-syndicate.org/commentary/ukraine-war-europe-energy-transition-fantasy-by-hans-werner-sinn-2022-11.

Smith, Adam. 1776. "Of Wages and Profit in the Different Employments of Labour and Stock." In *The Wealth of Nations*. Oxford, England: Bibliomania.com Ltd. Retrieved from the US Library of Congress, https://lccn.loc.gov/2002564559.

Timmer, Marcel P., Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J. de Vries. 2015. "An Illustrated User Guide to the World Input–Output Database: The Case of Global Automotive Production." *Review of International Economics* 23, no. 3: 575–605.

Velasco, Andrés, and Marcelo Tokman. 2022. "How to Get by without Russian Gas." Project Syndicate, April 28. https://www.project-syndicate.org/commentary/russian-gas-chiles-lessons-for-germany-europe-by-andres-velasco-and-marcelo-tokman-2022-04?barrier=accesspaylog.

Vogel, Lukas, Malte Neumann, and Stefan Linz. 2023. "Calculation and Development of the New Production Index for Energy-Intensive Industrial Branches." *WISTA* 2: 39–48.

Comments and Discussion

COMMENT BY

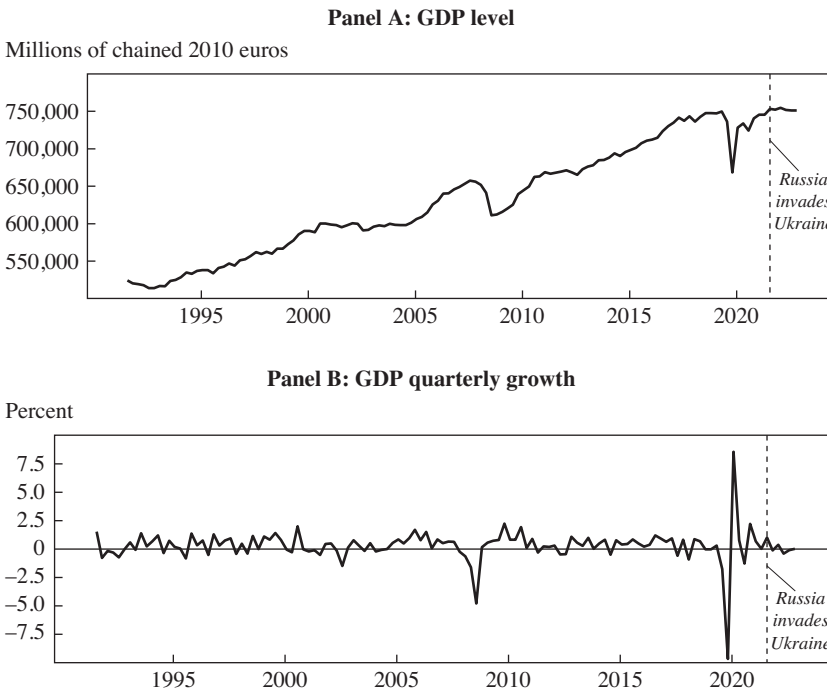
JAMES D. HAMILTON[1] Moll, Schularick, Zachmann, and their colleagues staked out a bold position in March 2022, predicting that loss of Russian natural gas would cause substantial but manageable challenges for the German economy (Bachmann and others 2022). They took a lot of flak for that conclusion from analysts who thought the economic consequences would be much more dire. But the subsequent events proved their prediction to have been largely correct. It's very appropriate at this point to provide a retrospective on how events unfolded a year and a half after Russia invaded Ukraine. I see my role as a discussant to be to highlight a number of the points made by Moll, Schularick, and Zachmann, perhaps with a slightly different emphasis from theirs.

ALL IS NOT WELL IN GERMANY The first point that bears repeating is that the German economy is currently struggling. Some in the financial press have started again referring to Germany as the "sick man of Europe" (*Economist* 2023). Panel A of figure 1 plots the level of German real GDP. Apart from the sharp drop and rebound associated with the COVID-19 pandemic, German output has essentially stagnated since 2019 and fell on average since the invasion.

Other measures corroborate that assessment. Panel A of figure 2 plots the Bundesbank's weekly index of the German real economic activity. This characterizes the German economy over the last year as experiencing a modest but clear decline. Panel B plots the ifo sentiment index based on a survey of German firms. Undeniably, many people in Germany have been very pessimistic about the economy since the invasion.

1. I thank Christiane Baumeister for assistance with obtaining the data for this discussion.

Figure 1. Level and Quarterly Growth Rate of German Real GDP

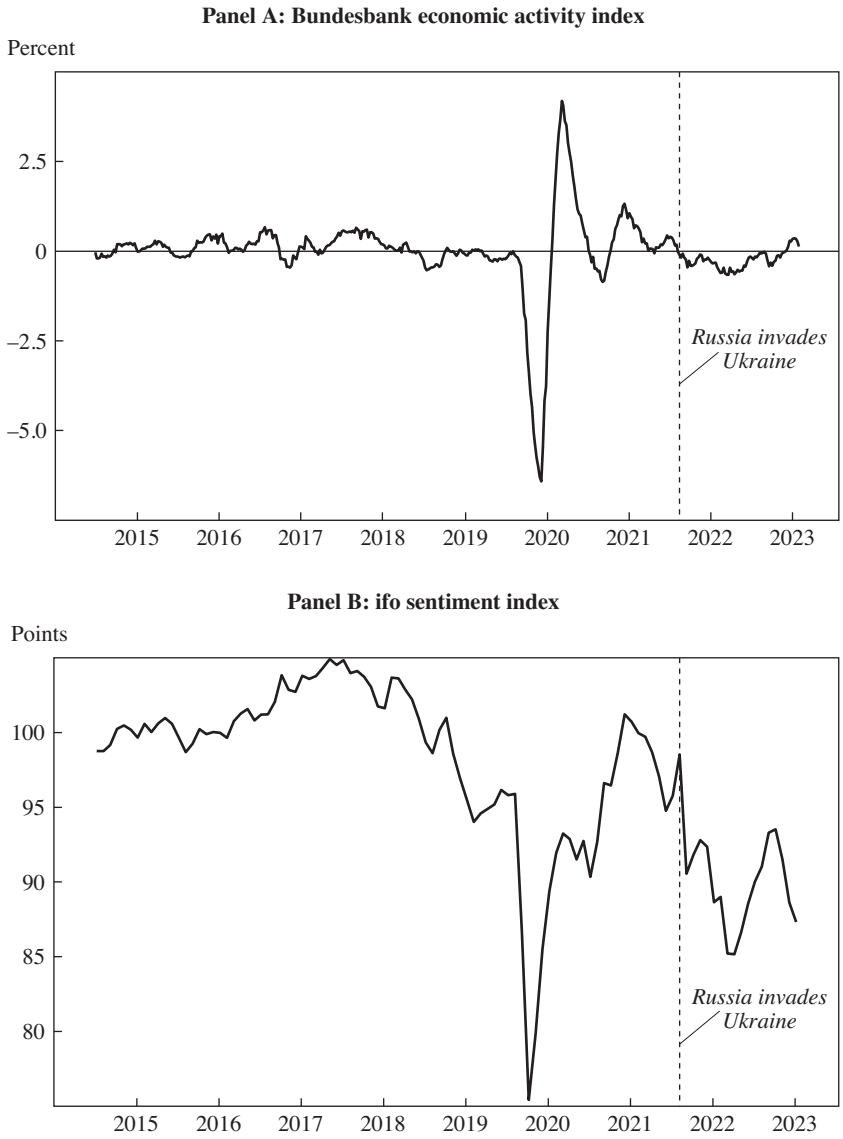


Source: Eurostat, series CLVMNACSCAB1GQDE, retrieved from FRED.
Note: Data are quarterly from 1992:Q2 to 2023:Q2.

To be sure, the challenges for the German economy began well before Russia invaded Ukraine. And the magnitude of the drop in output in 2022 is a far cry from the dire warnings of some prognosticators, and quite consistent with Bachmann and others (2022)'s original assessment of a substantial but manageable downturn. Still, I think we can agree that the German economy has faced some significant headwinds, and that disruptions in the supply of energy were part of those headwinds.

WHAT BROUGHT DEMAND DOWN? Figure 3 plots the wholesale price of natural gas in Germany. This exhibited a significant spike before the invasion, which Moll, Schularick, and Zachmann document was a result of prewar supply manipulations by Russia. The price went up spectacularly following the invasion. But natural gas prices began to fall dramatically after the summer of 2022 and are currently well below the levels even of 2021. Not only was the effect of the natural gas supply disruptions on German real

Figure 2. Other Measures of German Real Economic Activity



Source: Weekly Activity Index, Deutsche Bundesbank; and Business Climate Index for Germany, ifo Institute.

Note: Panel A data are weekly from January 5, 2015 to August 7, 2023. Panel B data are monthly from 2015:M1 to 2023:M7.

Figure 3. Wholesale Natural Gas Price in Germany

Source: Deutsche Börse Group.
Note: Wholesale price of natural gas in Germany, daily from January 2, 2013 to July 28, 2023.

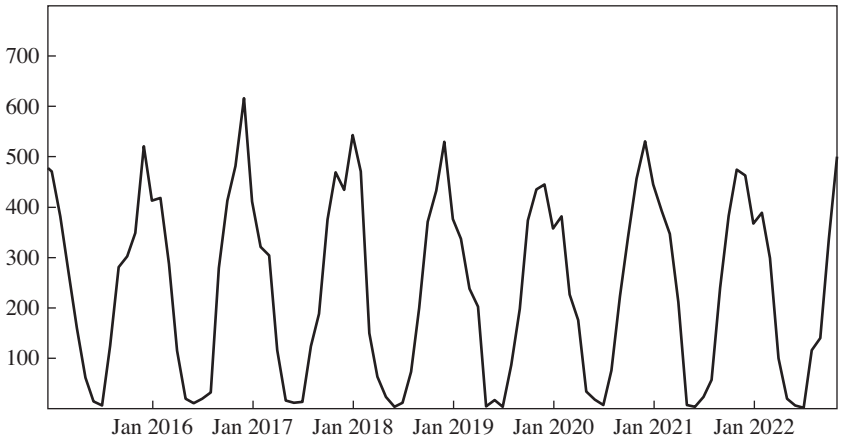
output more modest than many people had anticipated, so was the effect on the price of natural gas itself. One has to suspect that these two developments are related.

The first possibility many of us would consider is that there was some other factor shocking the demand, such as a milder than usual winter in 2023. But there's no real evidence that weather is the explanation (figure 4). The authors carefully investigate the contributions of weather to demand and conclude, correctly in my opinion, that weather is not the explanation for the mildness of the economic effects.

But why did the quantity demanded fall so much if the price actually fell? Part of the answer is the administered nature of the price paid by final users. This rose more slowly than the wholesale price, and the subsequent wholesale price declines were not immediately passed on to residential and business customers (Ruhnau and others 2023, fig. 1).

Another possible shift in the demand curve could arise from voluntary conservation efforts. The authors discount the importance of these, noting that the federal gas-saving campaign had a very limited budget. I would push back a little at the proposition that people only change their behavior if the government tells them to. I suspect that many German businesses and consumers felt a civic duty to conserve wherever they could. When the tanks

Figure 4. Heating Degree Days in Germany



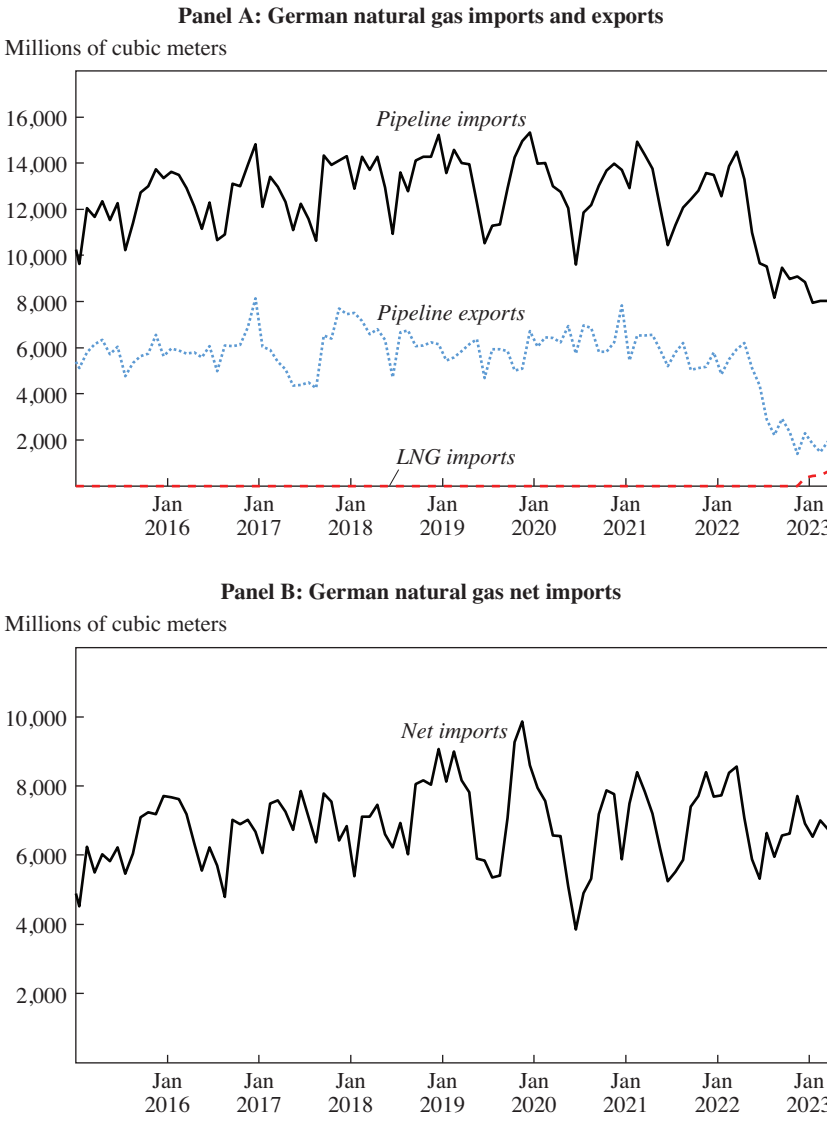
Source: Eurostat (data code nrg_chddr2_a).
Note: Heating degree days in Germany, monthly from 2015:M1 to 2022:M11.

are rolling into formerly peaceful villages, that may motivate some people to act in a way that government-sponsored advertising and slogans could not. I suspect that voluntary conservation may have played a role both in mitigating the price effects and, as I will elaborate below, in mitigating the real output effects as well.

Figure 5 highlights what I see as the single most important reason why reduced natural gas imports were less disruptive to the German economy than some had feared. The authors invested considerable effort into tracking flows of natural gas into and out of Germany. I have done something much simpler based on the gross flows reported in the Joint Organisations Data Initiative (JODI) database.[2] The top line in panel A shows that the monthly pipeline imports of natural gas into Germany fell by about 6 billion cubic meters, equivalent to 63 TWh per month and more than a 40 percent drop from preinvasion levels. Part of the initial worry came from people wondering: how in the world could Germany cut its use of natural gas by that much? The answer is, it didn't. As seen in the middle line in panel A of figure 5, most of the adjustment came in the form of reduced exports of natural gas from Germany. The loss of German net imports (panel B) is much more modest, around 2 billion cubic meters or

2. JodiGas, "The JODI Gas World Database," https://www.jodidata.org/gas/.

Figure 5. German Natural Gas Imports, Exports, and Net Imports



Source: JODI Gas World Database.

Note: Panel A presents data on German pipeline imports, LNG imports, and pipeline exports of natural gas. Data are monthly from January 2015 to May 2023. Panel B reflects total imports minus total exports.

21 TWh per month. This quick estimate is consistent with the cumulative decline in German consumption of 157 TWh that the authors arrived at in table 2 in the paper, using much more careful methods.

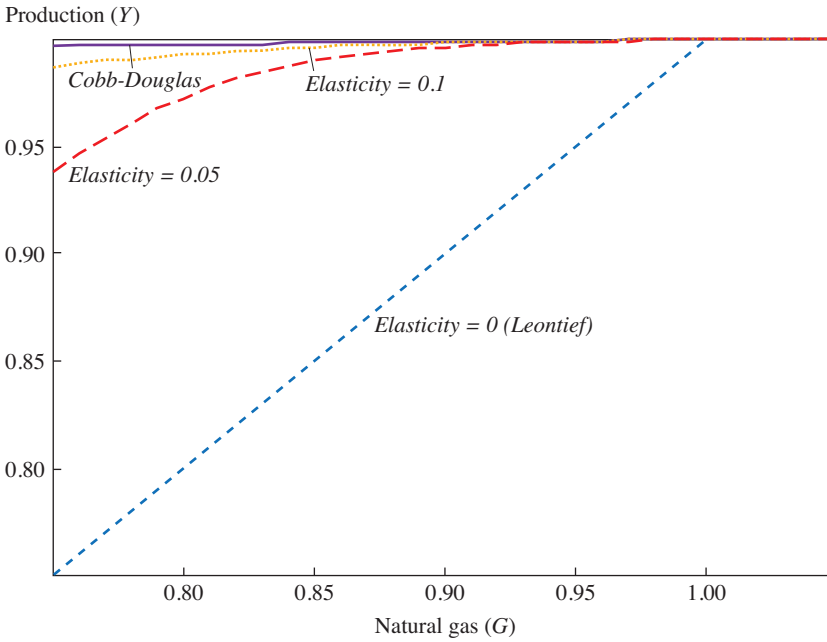
My conclusion is that the single biggest reason that the disruptions were less damaging to the German economy than some had feared is that Germany did not have to make the adjustments by itself. I see this very much as an illustration of the main point that the authors are making about the power of substitution. Markets find ways to adapt to challenges that policymakers and individual business planners can easily overlook. Their paper provides a wonderful demonstration of how this theme plays out in so many different ways.

DISCUSSION OF THE EFFECTS ON REAL GDP Let me now turn to the central question of the effects on overall real economic activity. Moll, Schularick, and Zachmann provide a simple illustration of the economy's ability to adapt using an aggregate CES production function:

$$Y = \left\{ \alpha^{1/\sigma} G^{(\sigma-1)/\sigma} + \left(1-\alpha\right)^{1/\sigma} \left[F\left(K, N\right) \right]^{(\sigma-1)/\sigma} \right\}^{\sigma/(\sigma-1)}.$$

Here Y is total real output, and G, K, and N are utilization of natural gas, capital, and labor, respectively, while σ is the elasticity of substitution and α determines the euro value of natural gas expenditures as a share of total nominal output. I take the initial expenditure share to be 1 percent for the calculations below. This corresponds to the authors' equation (1), where the only change I have made is to spell out explicitly the factors labeled as other inputs X in the authors' formulation. The question they ask is: what would happen to total output if utilization of natural gas were to change with utilization of capital and labor constant? The answer is plotted in figure 6, which reproduces the authors' figure 3. The graph shows how much Y would go down according to the above equation if natural gas consumption was cut by up to 25 percent while K and N did not change. If there is zero elasticity of substitution (corresponding to a Leontief production function), output would fall by the amount that natural gas was reduced. The authors' point is that the substitution elasticity can be very small, but as long as it is non-zero, effects are much more modest than would be predicted in the extreme Leontief case. For example, if $\sigma = 0.05$, output would only fall by 6 percent when consumption of natural gas is reduced by 25 percent. The authors' actual quantitative analysis is based on a detailed model of industry interactions

Figure 6. Effects of Changing Utilization of Natural Gas When Utilization of Capital and Labor Are Unchanged for Different Elasticities of Substitution



Source: Reproduced from figure 3 in the paper.
Note: Horizontal axis shows utilization of natural gas as a fraction of original level. Vertical axis shows total production as a fraction of original level.

as in Baqaee and Farhi (2019). But the simple summary in equation (1) gives some insight into what lies behind these calculations.

I have reproduced here the calculation in their figure 3 in order to highlight the implicit assumption that the drop in natural gas consumption does not lead to any change in utilization of capital or labor. I would argue that the defining characteristic of an economic recession is a dramatic decline in the utilization of capital and labor. From this perspective, one might say that analysis like that in their figure 3 rules out the possibility of an economic recession by assumption.

Is there a reason to think that a disruption in energy supplies could result in underemployed labor and capital? I've argued that, historically, underemployed labor and capital were very important in understanding why some historical oil price shocks were followed by economic recessions in the United States. We often observe in those episodes that the oil price

increases were followed by substantial declines in spending on new cars and other items. Quantitatively, the decline in car production made a significant contribution to the total observed decline in GDP in these historical downturns (Hamilton 2009). One can make a case that this correlation is causal. For example, the decline is the biggest for the least fuel-efficient vehicles, with production of more fuel-efficient cars sometimes even rising.

The original analysis by Bachmann and others (2022) recognized the potential importance of this issue. But they argued that it need not overturn their analysis, to the extent that "fiscal and monetary policies cushion potential demand-side Keynesian effects" (Bachmann and others 2022, 3). As long as we are taking this opportunity to praise the many ways in which their original analysis got so many things right, we should perhaps acknowledge that this particular policy prescription was not among them. I think we would all agree today that more fiscal and monetary stimulus was not an option for Europe in 2022. Indeed, the consensus view of many today is that excessive stimulus in 2021–2022 in Europe, the United States, and much of the rest of the world was a key factor in the resurgence of inflation. I would further argue that additional stimulus was also not an option in responding to the oil shocks of 1974 or 1979, for the same reason.

The authors were correct that mainstream macroeconomic models assume that demand effects could be mitigated using appropriate Keynesian-type stimulus. But that is not my view. I maintain that recessions do not result from a mismatch between aggregate demand and an aggregate production function, but instead from a mismatch between the composition of demand and the specific goods to which specialized resources are dedicated in advance to produce. Workers and factories may be capable of producing a huge number of gas-guzzling sports utility vehicles. But if people no longer want to buy those, the result is inevitably going to be underutilized capital and labor, for which added monetary stimulus is not the solution. I show how demand spillovers operating through these factors can play out in a dynamic general-equilibrium setting in Hamilton (2023).

In the present paper, Moll, Schularick, and Zachmann investigate possible demand spillovers in more detail than in the original study. They conclude that in the case of Germany in 2022, the observed magnitude of demand spillovers was limited. I agree with their analysis, and I think it is related to the authors' broader theme of the power of substitution. When gasoline prices double, the short-run options for most consumers are limited. They go ahead and fill up the gas tank, whatever it costs, and cut spending someplace else. In my view, it was those other cuts in spending that were the main cause of the economic disruption associated with historical oil price

shocks. The authors do a wonderful job of documenting the rich variety of ways that firms can (and did) reduce their use of natural gas without significant disruptions in other spending. And individual consumers can (and did) reduce their use of natural gas by lowering the thermostat, perhaps spurred in part by civic conscientiousness, and again without disrupting other economic spending. I believe that the authors are also correct that another reason why natural gas disruptions may be less disruptive than some historical oil shocks is the fact that the expenditure share of natural gas is on the order of 1 percent, in contrast with a number like 4 percent for the economic value of refined petroleum products. In my opinion, these were the primary reasons why the significant disruptions in GDP that some analysts had feared never came to pass.

SUMMARY There is much to like about this paper. I hope it will end up becoming a classic case study in the theme posed by the paper's title—the power of substitution.

REFERENCES FOR THE HAMILTON COMMENT

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick. 2022. "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia." Policy Brief 28. Bonn: ECONtribute.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "Networks, Barriers, and Trade." Working Paper 26108. Cambridge, Mass.: National Bureau of Economic Research. https://www.nber.org/papers/w26108.

Economist. 2023. "Is Germany Once Again the Sick Man of Europe?" August 17. https://www.economist.com/leaders/2023/08/17/is-germany-once-again-the-sick-man-of-europe.

Hamilton, James D. 2009. "Causes and Consequences of the Oil Shock of 2007–08." *Brookings Papers on Economic Activity*, Spring, 215–61.

Hamilton, James D. 2023. "Supply, Demand, and Specialized Production." Working Paper. University of California, San Diego. https://econweb.ucsd.edu/~jhamilto/H1.pdf.

Ruhnau, Oliver, Clemens Stiewe, Jarusch Muessel, and Lion Hirth. 2023. "Natural Gas Savings in Germany during the 2022 Energy Crisis." *Nature Energy* 8: 621–28.

COMMENT BY

TAREK A. HASSAN The paper studies the adjustment of the German economy after Russia cut Germany off from gas supplies in the summer of 2022. The authors highlight three main findings. First, despite Germany's

notable dependence on Russian gas, the gas cutoff proved to be a manageable shock for German firms. Second, the impact of the shock was transient, and its effects were primarily concentrated within a handful of sectors heavily reliant on gas. Third, German firms effectively employed two primary strategies for adjustment: reducing gas consumption and seeking alternative gas suppliers. Combined with good policy, these measures were sufficient to prevent a recession following the gas cutoff.

The insights presented in this paper and its precursors (Bachmann and others 2022a and 2022b) are invaluable, providing both academic and practical contributions. The authors illuminated the implications of canonical economic theory and distilled them into actionable policy recommendations at a time when such guidance was urgently needed.

Before turning to my main comment, I would like to reiterate two important points. First, sound economic policy was pivotal in the relatively benign outcome of the gas crisis. The German government found creative ways to make transfers to gas consumers that preserved price signals and incentivized them to reduce gas consumption. These schemes allowed the economy to adapt quickly and flexibly. In a similar vein, it is important to recognize that predictions based on aggregate production functions, like the ones made in the present paper, hinge on the preservation of price signals and incentives. There were many examples of poor policy decisions during this period, such as price caps and rationing. They serve as stark reminders of how the situation could have deteriorated.

Second, interconnected European gas markets played a vital role in mitigating the impact of the crisis (Papież and others 2022). Investments made since 2015 to connect the German gas market with the rest of Europe proved to be prudent and averted a recession.

Having made these initial points, the majority of this comment will concentrate on the broader issue of assessing the economic impact of ongoing economic shocks. Policymakers are frequently confronted with the challenge of dealing with impending or unfolding shocks. The German gas crisis is just one example in a landscape that includes crises like the COVID-19 pandemic, Brexit, sovereign defaults, government shutdowns, wars, and the like. More often than not, these shocks must be assessed and reacted to long before concrete data become available.

Measuring exposure to such shocks can be a complex endeavor, often difficult to accomplish beforehand. The German gas crisis, in this respect, was a comparatively straightforward case, as historical data on sector-specific gas imports were readily available. However, for many other types of shocks, identifying the affected firms and sectors can be challenging, if not impossible.

Typically, the approach is to make predictions based on economic theory, make a call on who is likely to be affected, and then wait several months to validate these predictions with accounting data. This was precisely the process followed in the case of the present paper: the authors had to formulate predictions based on theory (Bachmann and others 2022b), offer policy recommendations (Bachmann and others 2022a), and subsequently wait nearly a year and a half to perform a postmortem analysis in the present paper.

In essence, we often find ourselves in the unenviable position of comprehending the economic consequences of shocks only after they have occurred, rendering proactive policymaking difficult.

In the sections below, I will argue that systematic analysis of corporate earnings calls can offer real-time, powerful insights for analyzing the economic impact of ongoing and anticipated shocks. By examining what executives communicate to their investors about the state of their firms and their expectations regarding the impact of a given crisis, we can expedite quantitative analysis, allowing for more timely reactions and sound policy advice. I will argue that text-based data from earnings calls therefore hold substantial value for macroeconomic analysis.

The following sections of this comment will revisit the main steps of the authors' analysis, relying exclusively on the data generated from earnings calls available in 2022. Through this approach, I aim to demonstrate that analyzing these earnings calls could have led to similar conclusions in near real time, eliminating the need to wait for accounting data to become available and providing an opportunity for proactive and effective policymaking.

MEASURING EXPOSURE TO THE GAS SHOCK Executives at thousands of listed firms in eighty-two countries hold quarterly English-language calls with their analysts and investors to discuss any major issues confronting their firms. These high-stakes conversations typically begin with a management presentation, followed by a Q&A session where executives respond to analysts' questions. Transcripts of these earnings calls are widely available. I source them from London Stock Exchange Group and analyze them using NL Analytics.

The approach developed in Hassan and others (2019, 2023a, 2023b) measures the exposure, risk, and sentiment firms associate with a given shock—in this case, the cutoff from Russian gas—by analyzing what call participants say about the shock on their firm's quarterly earnings call.

What do German firms say about how a potential Russian gas shutoff will affect them?

The first step of the analysis is to generate a set of keywords associated with discussion of gas supply. For example, we may start with *gas supply*,

gas availability, *gas shortage*, *gas pipeline*, *Nord Stream*, and so on. There are very good methods for doing this in a systematic way. Here, I follow the approach in Bloom and others (2021), where I use an embedding vector model trained on earnings calls like a custom-trained thesaurus to give suggestions for different phrases executives might use when discussing reliance on Russian gas. For each suggestion, I then read ten randomly sampled excerpts of text where the phrase is mentioned in earnings calls to minimize false positives.[1]

We then use these keywords to find the sentences where call participants talk about gas supply. A simple measure of the extent to which a given firm expects to be affected by a possible gas shutoff is then simply to measure the number of sentences call participants devote to the subject in that firm's earnings call in that quarter:

$$(1) \qquad GasExposure_{i,t} = \#\text{ Sentences that mention gas.}$$

The intuition is simply that managers and analysts devote more time to events of greater importance to the firm.

Second, to measure the amount of risk call participants associate with the shock, we count which of the sentences identified in equation (1) also mention *risk*, *uncertainty*, or any synonym thereof (Hassan and others 2019).[2]

$$(2) \qquad GasRisk_{i,t} = \#\text{ Sentences that mention gas and risk synonym.}$$

We may think of *GasRisk* as the second-moment impact of the shock—a measure of how much uncertainty it generates for the firm. Finally, to distinguish first-moment impacts (bad news) from the shock's effect on risk, it is sometimes useful to measure the sentiment with which call participants discuss the shock

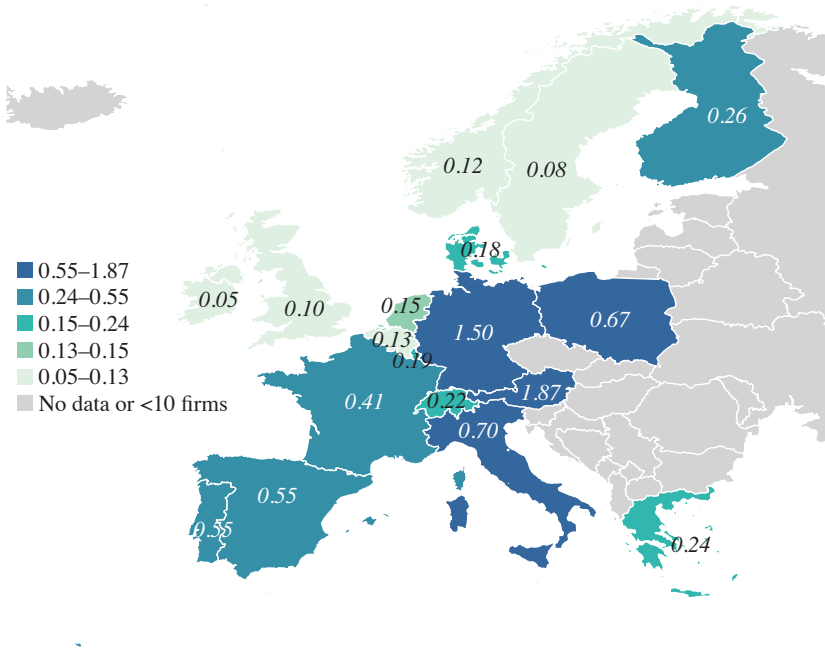
$$(3) \qquad GasSentiment_{i,t} = \#\text{ Positive} - \#\text{ Negative sentences that mention gas.}$$

Loughran and McDonald (2011) provide a widely used library of tone words to make this distinction.

1. For methods that do not require human intervention, see Hassan and others (2019) and Sautner and others (2023).

2. Single-word synonyms of *risk*, *risky*, *uncertain*, and *uncertainty* as given in the *Oxford Dictionary* (excluding *question* and *questions*).

Figure 1. Natural Gas Exposure in European Countries, 2022:Q3



Source: NL Analytics.

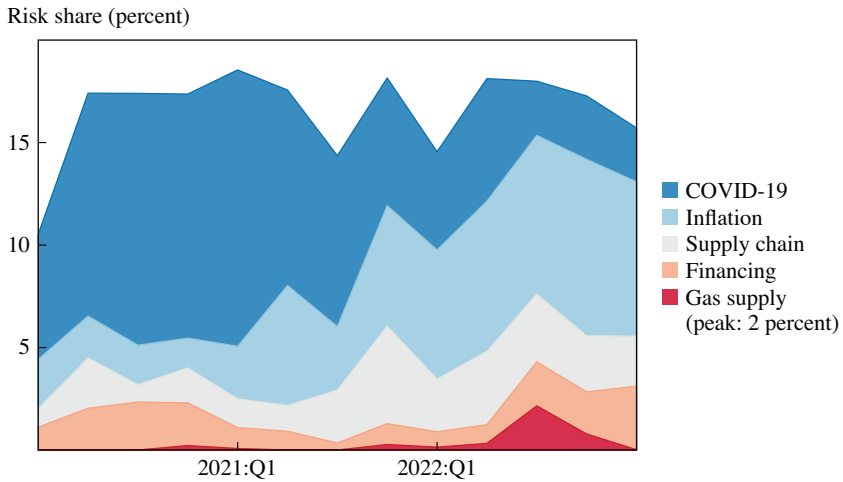
Note: This figure illustrates the spatial distribution of natural gas exposure across European countries. The numbers provided represent the average count of sentences that mention natural gas supply in all earnings calls held by firms headquartered in each respective country during the third quarter of 2022. Countries with fewer than ten transcripts in that quarter are excluded.

One useful feature of these data is that they are at the firm-quarter level, which can then be merged with conventional firm-level data from Compustat Global and other sources.

The third step is then to analyze the data. Importantly, because each of these series is generated from text, we can also use them to identify the most important pieces of text to read to understand the country-, firm-, and sector-level variation in our measures of gas exposure, risk, and sentiment. I will show one example of such targeted reading below.

Figure 1 shows the variation in *GasExposure* across European firms in the third quarter of 2022. Notably, it illustrates that German and Austrian firms exhibited the highest degree of exposure, dedicating a substantial portion of their discussions to this issue (mentioning the possibility of a Russian gas shutoff 1.87 and 1.5 times on average in their earnings calls).

Figure 2. Decomposition of Risks Discussed by German Companies, 2020:Q1–2023:Q1



Source: NL Analytics.

Note: The figure illustrates the proportion of risk mentions among 190 German firms related to gas supply in comparison to four other topics: COVID-19, inflation, supply chain disruptions, and financing challenges. The fraction for each topic is calculated by dividing the number of sentences that mention risks associated with that topic by the total number of sentences that reference risk in general.

Additionally, a striking geographic pattern emerges, as the countries with a relatively larger number of mentions cluster closely together. This spatial correlation underscores the regional nature of the impact, indicating a shared concern among Central and Eastern European nations about the possible gas cutoff.

GAS EXPOSURE VERSUS OTHER RISKS FACED BY GERMAN FIRMS Although it is clear that German firms were more concerned about gas supply than firms in many other European countries, it is important to know how the threat of a gas shutoff compares with other prevailing concerns at the time.

A useful way of making such a cardinal comparison between different types of risks is by considering what share of mentions of risks is attributable to gas supply relative to other topics. An average earnings call transcript tends to contain about six sentences that mention risk, uncertainty, or a synonym thereof. Figure 2 shows the composition of risk discussions among the 190 German firms in our sample. It shows what fraction of risk mentions corresponds to gas supply and each of four other topics: COVID-19, inflation, supply chain disruptions, and financing challenges.

In the early part of the sample, there was a pronounced anxiety tied to COVID-19, but by 2022:Q3, these fears had markedly diminished. The graph underscores that inflation-related risks have overshadowed other concerns since early 2022, emerging as the predominant risk for German firms (7.7 percent of all mentions of risk in the third quarter of 2022). Concurrently, worries related to the supply chain have also risen to prominence (3.3 percent). Notably, concerns related to both financing and the Russian gas supply stand at a relatively low 2.1 percent and 2.2 percent, respectively.

In other words, even at the height of the Russian gas crisis, concerns about gas supply are on par with or even secondary to a range of other concerns faced by the average German listed firm. Second, while concerns about inflation, supply chain, and COVID-19 are highly persistent, the anxiety around the gas supply sees a brief spike and then rapidly dissipates, contrasting sharply with the enduring concerns tied to other risk domains.

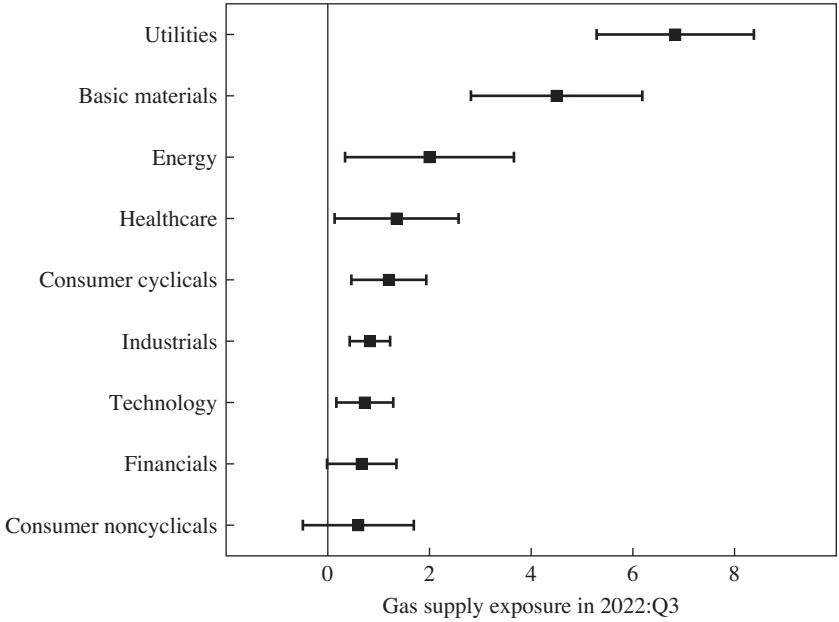
SECTOR-LEVEL IMPACT Although the Russian gas crisis was not the most urgent concern for the average German firm, it was a major source of concern for some firms in specific sectors. These, no doubt, were also highly vocal in the public discourse on the subject.

Figure 3 shows the average number of mentions of the Russian gas crisis across sectors. Evidently, the impact is highly concentrated. German utility companies, in particular, devoted significant attention to this issue, underlining its critical importance for them. Similarly, firms within the basic materials sector, which includes notable entities like BASF, exhibited significant conversations on the subject. Conversely, the remaining sectors exhibit much lower exposure. Consumer noncyclicals, for instance, only registered an average of 0.2 mentions of gas supply during the same period. Again, this evidence is consistent with the authors' assertion that a cascading failure of the German economy was never in the cards.

ADJUSTMENT TO THE SHOCK In a final step, I delve deeper into the authors' examination of how German firms adjusted to the challenges posed by the gas crisis. To accomplish this, I undertake a targeted reading of executives' statements regarding their plans. To this end, I download the text encompassing all 330 mentions of natural gas supply made by German firms from June 1, 2022, to December 31, 2022. Within this corpus, 157 sentences discuss specific strategies for addressing the crisis.

A rough reading of these text snippets reveals four primary categories of adjustment strategies embraced by German firms: a transition toward alternative energy sources, reductions in gas consumption, a shift toward

Figure 3. Average Mentions of Russian Gas Crisis across Sectors, 2022:Q3



Source: NL Analytics.

Note: This figure illustrates the average natural gas exposure of German firms across sectors in 2022:Q3. The exposure for each sector is determined by averaging mentions of natural gas supply across its firms for the quarter. Whiskers indicate 95 percent confidence intervals.

alternative suppliers of natural gas, and a reliance on government assistance. Table 1 gives examples of executives' statements in each category.

Figure 4 depicts the proportion of text excerpts referencing each of the four mitigation strategies. Switching to alternative fuels, such as oil or electricity, accounts for 30 percent of the mentions. Measures centered around curbing gas consumption comprise 25 percent; 23 percent discuss the identification of alternative gas suppliers, while 8 percent mention strategies that hinge on obtaining government assistance.

Interestingly, despite stemming from distinct data sources, these observations align seamlessly with the authors' conclusions. Both sets of findings underscore the significance of demand reduction and the pursuit of alternative gas sources as primary mechanisms that curtailed a larger impact of the gas crisis on the German economy.

STATEMENTS IN EARNINGS CALLS VERSUS THE MEDIA Before concluding, it is worth highlighting the differing communication styles executives choose

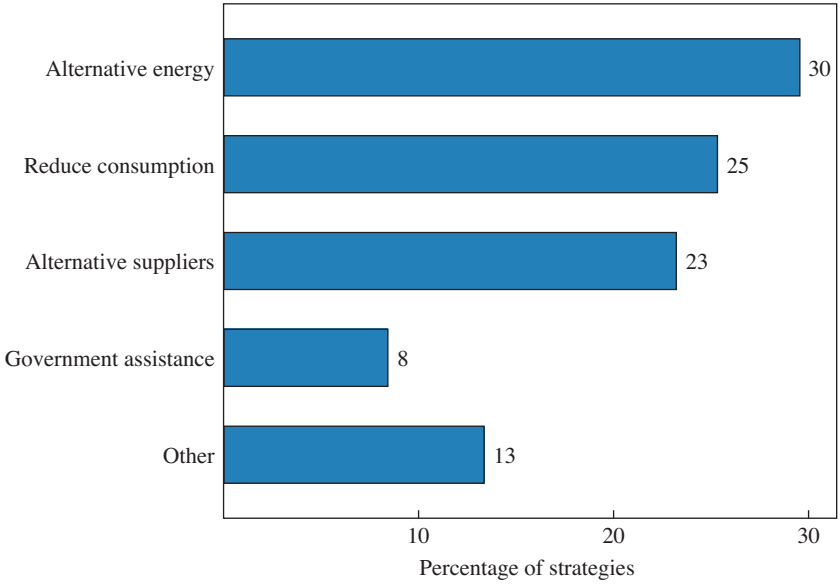
Table 1. Firm Strategies for Adjusting to Gas Cutoff

Strategies	*Transcript excerpts*
Alternative energy	"We can generate steam which we need for our production with fuel oil, electricity instead of natural gas." (Aurubis AG, Mineral Resources, August 5, 2022)
	"If needed, we are able to switch the heating supply and that's mainly for the painting from gas to heating oil in the short term and that's 100 percent." (Deutz AG, Industrial Goods, August 11, 2022)
Reduced consumption	"We prepared ourselves since the beginning of the war on the Ukraine to reduce our gas consumption as best as possible." (Infineon Technologies AG, Technology Equipment, August 3, 2022)
	"But overall, this is an expression of the fact that we've actually consumed significantly less gas. And if your question is how much less is in the order of magnitude of almost 40 percent lower gas consumption in Q3 than in the prior year quarter." (BASF SE, Chemicals, October 26, 2022)
Alternative suppliers	"We are also helping to diversify gas supply in Europe through investments in LNG infrastructure and LNG imports." (RWE AG, Utilities, August 11, 2022)
	"The further diversification of our gas procurement is well on track." (EnBW Energie Baden Wuerttemberg AG, Utilities, August 12, 2022)
Government assistance	"Since the beginning of the war, there have been regular meetings between German industry and the German government to look at scenarios for gas and other things." (Mercedes Benz Group AG, Automobiles & Auto Parts, July 27, 2022)
	"We are aware of and accept our responsibility for the health of millions of people. As such, we are confident in being granted priority access to gas supplies in the event of restrictions." (Gerresheimer AG, Healthcare Services & Equipment, July 13, 2022)
Other	"We have already significantly reduced our exposure in Germany by implementing preemptive measures. These include rearranging gas consumption between sites." (Beiersdorf AG, Personal & Household Products & Services, August 4, 2022)

Source: London Stock Exchange Group.

when addressing the public versus their investors. Those individuals who might have been motivated to amplify the projected effects of the Russian gas cutoff on their businesses in public media statements often conveyed a more balanced perspective during their earnings calls. For instance, Martin Brudermüller, the leader of BASF, mentioned in a newspaper interview dated March 31, 2022, that the cessation of Russian gas "could bring the German economy into its worst crisis since the end of World War II and destroy our prosperity" (Brankovic and Theurer 2022). Yet, in the earnings call on April 29, he detailed BASF's strategic response to reduced gas consumption, stating: ". . . [W]e have increased and will further increase our sales prices to pass on higher natural gas prices. At our European sites,

Figure 4. German Firms' Strategies for Adjusting to the Gas Crisis, 2022:Q3 and 2022:Q4



Source: NL Analytics.

Note: Percentages for each strategy were calculated by dividing the number of mentions for a specific strategy by the total number of strategy mentions (157) during that period.

where technically feasible, preparations to substitute natural gas by alternative feedstocks are ongoing" (BASF 2022a, 5). Furthermore, the management team acknowledged their capability to decrease gas usage by up to 50 percent without ceasing production (BASF 2022b).

CONCLUSION Policymakers are frequently tasked with addressing the implications of sudden economic shocks. The Russian gas shutoff serves as a quintessential example of such a shock.

In this comment, I have posited that a systematic analysis of earnings calls offers a powerful lens to understand and quantify the impending and immediate impact of such shocks in near real time and before conventional data sources are available. Doing so can provide policymakers with timely data pivotal for shaping policy decisions.

Examining earnings calls held by German and European firms in 2022 fundamentally confirmed the authors' conclusions. We found that German industry was exceptionally dependent on Russian gas. Yet, despite this dependence, the Russian gas shutoff represented a surmountable challenge for German firms—on par with concerns about supply chains and financing

constraints but less concerning than, for example, the historically high levels of inflation that prevailed at the time. The gas shock was transitory and highly localized, with its effects predominantly felt within the utilities and basic materials sectors, with no evidence of cascading failures in other sectors. When navigating this episode, the predominant strategies employed by German firms revolved around curtailing their consumption of gas, substituting other sources of fuel, and switching to alternative gas suppliers.

In sum, the insights from my text-centric evaluation align seamlessly with the authors' conclusions, underscoring the validity of their results.

REFERENCES FOR THE HASSAN COMMENT

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Ben McWilliams, and others. 2022a. "How It Can Be Done." Policy Brief 34. Bonn: ECONtribute.

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick. 2022b. "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia." Policy Brief 28. Bonn: ECONtribute.

BASF. 2022a. "Analyst Conference Call Q1 2022." April 29. Speech. https://www.basf.com/global/documents/en/investor-relations/calendar-and-publications/presentations/2022/BASF_Speech_Analyst_Conference_Call_q1-2022.pdf.assetdownload.pdf.

BASF. 2022b. "Conference Call Q1 2022 Transcript Q&A." April 29. Transcript. https://www.basf.com/global/documents/en/investor-relations/calendar-and-publications/presentations/2022/BASF_Q1-2022_Transcript_QA_by_Topic.pdf.assetdownload.pdf.

Bloom, Nicholas, Tarek A. Hassan, Aakash Kalyani, Josh Lerner, and Ahmed Tahoun. 2021. "The Diffusion of New Technologies." Working Paper 28999. Cambridge, Mass.: National Bureau of Economic Research. http://www.nber.org/papers/w28999.

Brankovic, Maja, and Marcus Theurer. 2022. "BASF Chef im Interview: 'Wollen wir sehenden Auges unsere gesamte Volkswirtschaft zerstören?'" [BASF boss in an interview: "Do we want to destroy our entire national economy?"] *Frankfurter Allgemeine Zeitung*. March 31. https://www.faz.net/aktuell/wirtschaft/unternehmen/basf-chef-warnt-vor-gas-embargo-schaeden-fuer-deutsche-volkswirtschaft-17925528.html.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun. 2023a. "Firm-Level Exposure to Epidemic Diseases: COVID-19, SARS, and H1N1." *Review of Financial Studies* 36, no. 12: 4919–64.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2019. "Firm-Level Political Risk: Measurement and Effects." *Quarterly Journal of Economics* 134, no. 4: 2135–202.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2023b. "The Global Impact of Brexit Uncertainty." *Journal of Finance* 79, no. 1: 413–58.

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66, no. 1: 35–65.

Papież, Monika, Michał Rubaszek, Karol Szafranek, and Sławomir Śmiech. 2022. "Are European Natural Gas Markets Connected? A Time-Varying Spillovers Analysis." *Resources Policy* 79:103029.

Sautner, Zacharias, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang. 2023. "Firm-Level Climate Change Exposure." *Journal of Finance* 78, no. 3: 1449–98.

GENERAL DISCUSSION David Romer said that concrete examples of substitution, such as BASF's shift from domestically produced ammonia to imports from its American plants, are helpful in showing how the paper's main findings manifested themselves in practice. Romer also suggested that the authors avoid using the phrase "decline in demand" to refer to a fall in the quantity demanded. He commented that authors don't currently seem to answer the question of whether the reduction in household gas consumption corresponded to a movement along or a shift of the demand curve. On the one hand, when faced with higher prices, consumers may reduce their consumption or invest in more energy-efficient products—a movement along the demand curve. On the other hand, along the lines of James Hamilton's discussant remarks, if consumers were aware of a possible national crisis and reduced their gas consumption in response, that would constitute a shift of the demand curve.

Caroline Hoxby noted that there were large reductions in household gas consumption, despite gas prices rising only marginally. Hoxby argued that households typically have fewer substitution alternatives relative to firms, and she inquired about the forms of substitution that households might have engaged in, such as adjusting thermostats or investing in warmer clothing.

Claudia Sahm highlighted that the reduction in household gas consumption mirrored the reduction in gas demand by firms. She pointed out that businesses faced the market price for gas, whereas households benefited from incentives to reduce consumption and price caps on natural gas use (Gaspreisbremse). Sahm concluded that firms and households have different substitution possibilities and was intrigued by their almost equivalent reduction in gas demand. On a prior visit to Germany, she observed the government advertisements urging reduced gas use, which she believed supported Hamilton's assessment that a portion of the household consumption drop

can be attributed to an emotional response to the Russian invasion of Ukraine and a desire to assist in national efforts.

Moritz Schularick responded that households adopted measures to reduce the quantity of gas demanded in response to higher prices, such as lowering their thermostats, refraining from heating unused rooms, and sealing their doors to prevent cold drafts. However, households also acted to make substitutions for their gas consumption, such as purchasing pellet ovens. Schularick noted that, in most cases, households respond to price changes, but the exact dynamics are complex because retail prices are reset annually and many gas providers operate under longer-duration contracts. Many gas providers failed because they were obligated to provide gas at a low price stipulated by the contracts. The German government intervened to reset some of these contracts. Schularick agreed with Sahm's comment that households likely took efforts to restrict their gas usage for the sake of national security.

Angelos Theodorakopoulos pondered the evidence that German firms replaced significant amount of Russian gas with imports from elsewhere. If such third countries are also highly reliant on Russian gas, the apparent decoupling between overall industrial production and gas-intensive production may actually be trade diversion. He also commented that the aggregate production function modeled in the paper assumes a large cost share for material inputs, which, when produced domestically, are reliant on the Russian gas products. Theodorakopoulos argued that true decoupling does not occur if these material inputs are outsourced from countries that are also highly exposed to Russian gas and oil. Decoupling is persistent, he stated, as opposed to firms and households reducing consumption, relying on stockpiles, and subsequently reverting to their previous behavior.

Georg Zachmann commented that the authors had conducted an analysis that modeled the European gas system as an input-output matrix, allowing them to identify how reduced gas demand in neighboring countries contributed to German supplies. A drop in the Dutch gas demand accounted for 4 percent of German gas supply in the past year, with Belgium contributing 1 percent and France 0.5 percent. Zachmann concluded that the reduced demand in adjacent countries had a substantial impact on German gas supply. He also highlighted that the resilience of the European internal market following the cutoff from Russian gas played a substantial role in Germany's economic adaptability. The country would have suffered greatly on its own, he hypothesized.

Elaine Buckberg noted that the large reduction in gas demand from both producers and households could have potential climate implications. She

posited that some of the observed industrial compression was likely due to geographical shifts in production. However, compression not attributable to the relocation of production to lower-cost countries could be of interest from a climate standpoint, she argued. Buckberg wondered whether declines in gas usage could become permanent.

Benjamin Moll responded that the German gas cutoff could potentially be used as a natural experiment to estimate elasticities and examine links in the supply chain. He lamented the limited availability of data but explained that the German statistical agency would be releasing information on gas usage at the sector level later in the year. He advocated for more climate research using data from the gas crisis.

In response to Hamilton's discussant remarks, Schularick agreed that the reactions of firms and households to the German gas cutoff were dependent on effective price signals. He emphasized that agents respond to incentives, linking this idea to the broader discussions on climate change and substitution. The authors investigated the distributional consequences of the gas cutoff and found no evidence that its impacts were regressive. Zachmann emphasized that data from the German gas crisis would be useful in predicting the response of household demand to various shocks and the ability of certain sectors to adapt. Elasticities estimated using these data, he added, could be useful in determining the economic consequences of decarbonization.

Alan Blinder inquired about the significance of liquified natural gas (LNG) and remarked that while the construction of terminals and the processing of natural gas are considered time-consuming, LNG import capacity was critical in the wake of the gas cutoff.

Randall Kroszner noted that substitution could be applied to climate change issues. In response to Tarek Hassan's discussant remarks, Kroszner remarked that analyzing earnings calls to identify differences between what companies announce publicly and what they communicate to investors could inform policy. He also contended that public policy could ease the frictions associated with substitution. Kroszner noted that in Germany, for example, regulators expedited the installment of new LNG terminals, a process that would usually require many layers of approval. In May 2022, the German government approved the terminals, and they were operational in December of the same year, Kroszner explained. He concluded that substitution could take place more easily with adaptive regulation.

Schularick explained that the German LNG terminals were constructed quickly as a policy response to the cutoff from Russian gas. However, a large portion of the LNG imported to Germany came through existing

ports and via the Netherlands. He remarked that the floating LNG ports built in the wake of the cutoff, although currently operational, were small and made minimal contributions to the adaptability of German economy. Zachmann agreed, noting that the German LNG terminals became operational in early 2023, and thus they did not play a decisive role in Germany's response to the limitations on Russian gas imports. Currently, the LNG terminals mitigate supply constraints. He credited the moderate economic impacts of the cutoff to demand reductions in neighboring countries and their willingness to supplement limited German supply.

Rebecca Freeman commented that the moderate reductions in output following Germany's cutoff from Russian gas might be a unique outcome, pointing to the particular adaptability of the German gas supply chain, which was able to transition to alternative gas suppliers relatively easily. Freeman pondered whether sectors with more complex or vulnerable supply networks would respond similarly.

Benjamin Golub emphasized that government coordination amid supply chain disruptions is critical to facilitating substitution. Golub also noted that some supply shocks result in smooth adjustments, while others generate an abrupt, discontinuous response. There are examples of both outcomes in complex systems, Golub explained.

Benjamin Harris brought up the supply shocks to semiconductor production, which had an impact on the manufacturing of electronics and automobiles and led to inflation. He pondered why the supply shock to German gas, precipitated by limited Russian imports, generated only marginal effects on production, while the supply shock to semiconductors had more substantial economic consequences. Harris suggested an analysis of the recent supply shocks and their varying impacts on economic activity.

Yongseok Shin commented on the paper's potential climate change implications. A possible unintended takeaway, Shin noted, is that substitutability will allow the economy to adjust easily in the face of climate disasters. He added that with ample planning time, it is possible to determine the most effective ways to substitute.

George Akerlof turned the attention to what he termed an opposite shock—one to the global food supply. Wealthy countries may only face moderate economic effects, while the resultant price increases in poor countries would mean that the population cannot afford food, Akerlof explained. He added that in low-income countries, food will be exported to rich nations, rather than feeding the domestic population. Akerlof noted that such dynamics could be precipitated by global warming, which could produce such shocks to the food supply.

Angus Deaton emphasized that the predictions about German production by economists based on theories of substitutability had proved more accurate than industry analyses. He noted that it is tempting to treat such successes as demonstrations of the superiority of economic tools, but economists need to avoid professional hubris and remember those many occasions on which they were very wrong.

In response to Hassan's discussion, Jason Furman offered an anecdote: in a meeting regarding the 2014 Crimean crisis, the CEO of one of the top five largest oil companies in the world told Furman that the American sanctions on Russia would destroy the company and American jobs. A week later during an earnings call, the CEO assured investors that the sanctions would have no effect on the company—a testament to Hassan's discussion, Furman added.

Furman further commented that, except for short-run demand policy, rigorous modeling of economic phenomena that are commonly discussed in the public domain often reveals minimal impacts: for example, macroeconomic analysis of subjects such as the trade war with China, new trade agreements, infrastructure plans, childcare, and tax reforms, show changes that are mere basis points of annual growth rates. Furman questioned whether the real-world impacts are genuinely as small as these models predict, or if models are instead missing important components, meaning the actual economic implications of policy and macro phenomena are much larger than estimated.

Schularick mentioned his interest in conducting additional analysis using data from earnings calls to identify systematic differences between companies' public announcements and their communications to investors. He recalled that firms from an array of industries voiced concerns over the economic ramifications of restrictions on Russian gas. Schularick noted the automobile industry in particular, which expressed concern about the cutoff publicly but scarcely addressed it during earnings calls.

Justin Wolfers noted the complexity of constant elasticity of substitution models. He proposed an exercise that ranks sectors by energy intensity and simulates a shutdown of the most energy-intensive industries until the quantity of gas demanded declines by 20 percent. Wolfers argued that this type of simulation can be used to gauge the effects of an energy supply shock on the German economy and may be a more effective way of communicating findings.

Zachmann explained that by focusing solely on the most energy-intensive sectors, one could easily observe a 20 percent decline in demand. The most affected sectors are capital- and energy-intensive and have low employment

and value-added. Zachmann pondered potential strategies for these industries, such as allowing market forces to determine their fate or providing subsidies in the hope that energy will be cheaper in the future. Şebnem Kalemli-Özcan brought up one of her own papers, which calibrates an open economy adaptation of the Baqaee-Farhi model that considers both trade and domestic elasticities.[1] She emphasized the importance of differentiating short- and long-run elasticities and explained that the most significant source of variation in predicting the price impact of a supply shock is the shift in the elasticity of substitution above and below one. Using domestic and international elasticities calculated by Boehm, Pandalai-Nayar, and Levchenko,[2] the model highlights the differing price impacts of a supply shock in the near and far term as elasticities shift to imply substitutes rather than complements in production, Kalemli-Özcan explained. She argued that this type of model can accurately explain price and output dynamics.

In response to Hamilton's discussion on the business cycle effects of the German gas cutoff, Schularick offered several comments. He explained that the paper did not include a full analysis of the business cycle effects due to time constraints. Schularick elaborated on the findings of the Baqaee-Farhi model presented in the paper. Specifically, the model predicted a GDP decline of about 1 percent but acknowledged an upper bound of about 3 percent, accounting for the business cycle amplification effects. Some of the authors of the original "what if?" paper[3] produced a subsequent publication with a more comprehensive analysis of the business cycle implications. They found similar effects when allowing for business cycle amplification.

1. Julian di Giovanni, Şebnem Kalemli-Özcan, Alvaro Silva, and Muhammed A. Yildirim, "Pandemic-Era Inflation Drivers and Global Spillovers," working paper 31887 (Cambridge, Mass.: National Bureau of Economic Research, 2023), https://www.nber.org/papers/w31887.

2. Christoph E. Boehm, Andrei A. Levchenko, and Nitya Pandalai-Nayar, "The Long and Short (Run) of Trade Elasticities," *American Economic Review* 113, no. 4 (2023): 861–905.

3. Rüdiger Bachmann, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick, "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia," policy brief 28 (Bonn: ECONtribute, 2022).