The Journal of

Economic Perspectives

A journal of the
American Economic Association

Summer 2021

The Journal of
Economic Perspectives

A journal of the American Economic Association

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

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

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

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2021, Vol. 35, No. 3. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $24.00, $34.00, or $44.00 depending on income; for an additional $15.00, you can receive this journal in print. The journal is freely available online. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 35 • Number 3 • Summer 2021

Symposia

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 3–24

Effects of the COVID-19 Recession on the US Labor Market: Occupation, Family, and Gender

Stefania Albanesi and Jiyeon Kim

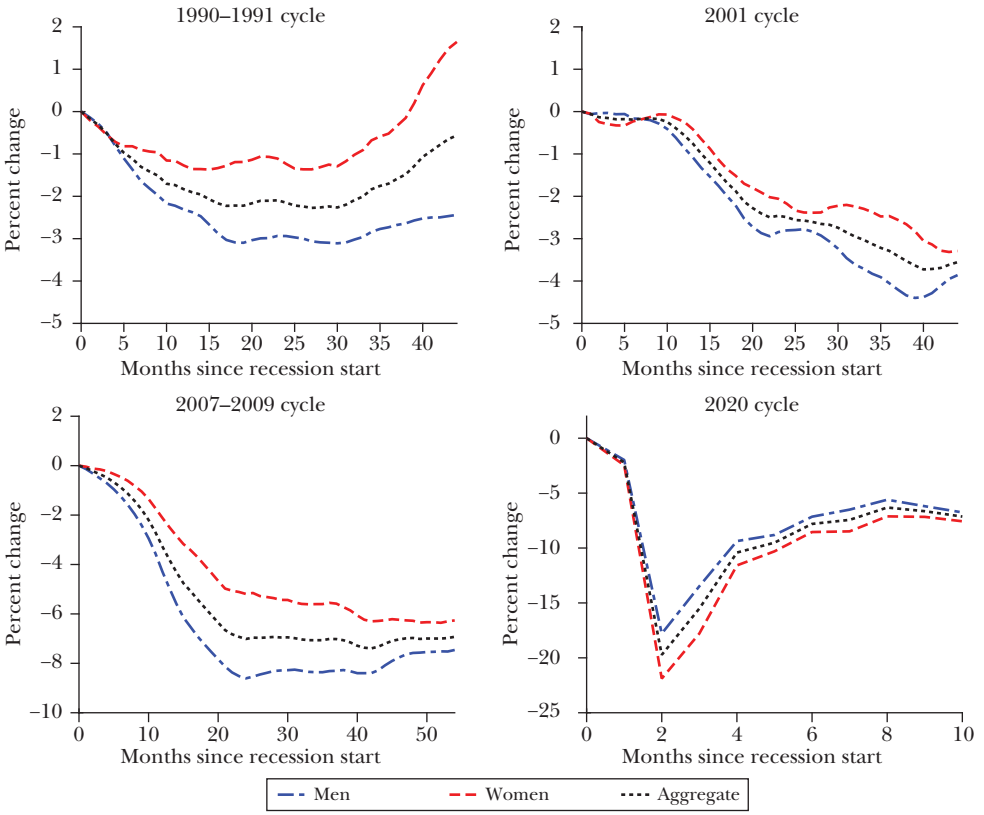
Recessions in the United States are usually associated with a larger employment drop for men than for women. But during the COVID-19 recession, employment losses were larger for women. Figure 1 shows the employment-to-population ratio for men and women during the last four business cycles. The drop in the ratio was higher for men than for women in each previous cycle, but not in the pandemic recession.

There are demand-side and supply-side reasons why the pattern of employment changes during recessions is different for men and women, and these patterns have not been the same during the pandemic as in previous recessions. On the demand side, the asymmetry is partly explained by gender differences in the occupation distribution, with men primarily employed in production occupations and women concentrated in service occupations, which tend to be less cyclical (Albanesi and Şahin 2018; Olsson 2019). During the pandemic, however, there has been a sizable drop in the demand for services as a result of both the mitigation measures enacted to contain the pandemic and consumers' response to the risk of infection (Chetty et al. 2020). Given the concentration of women in service occupations, they have been disproportionately hit by the corresponding employment losses. On the supply side, married women have, in the past, tended to increase their attachment to the

■ *Stefania Albanesi is Professor of Economics, University of Pittsburgh, Pittsburgh, Pennsylvania. She is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and a Research Fellow, Centre for Economic Policy Research, London, United Kingdom. Jiyeon Kim is an Associate Fellow, Korea Development Institute, Yeongi-gun, South Korea. Their email addresses are stefania.albanesi@gmail.com and jik51@kdi.re.kr.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.3.

Figure 1
Percentage Change in the Employment-to-Population Ratio since the Start of Each Recession for the Four Most Recent Business Cycles



Source: Authors' calculations based on Current Population Survey.
Note: Recession dates based on the National Bureau of Economic Research business cycle dates.

labor force during economic downturns relative to expansions as a form of household insurance that reduces the impact of recessions (Ellieroth 2019). Before the pandemic, the lower cyclicality of women's employment led to a reduction in the cyclical volatility of aggregate employment as the share of women in the workforce increased from the 1970s onward (Albanesi 2019). During the pandemic, limited availability of in-person childcare and schooling options led many parents—and women in particular—to exit the labor force.

In this essay, we first focus on the differences in supply-side employment responses of men and women during business cycles, in part using a comparison between the Great Recession and the pandemic recession to illustrate. We then turn to occupational differences and how they influenced employment for men

and women using monthly data in 2000. To do so, we classify occupations by their exposure to the pandemic, based on contact intensity and ability to work remotely and show that women are overrepresented in high-contact and inflexible occupations most affected by the pandemic.

We then explore the relative importance of the supply-side and demand-side responses in two ways. First, we use a regression approach to analyze the employment changes of women and men during the pandemic. We focus on differences in family status but also show that controlling for occupations attenuates the decline in employment and the gender differences by about one-third. We then look at gross flows of labor. For example, the flow from employment to nonparticipation can be viewed as a supply-side withdrawal from the labor market, while the flow from employment to unemployment can be viewed as driven from the demand-side of the labor market. We find that employment to nonparticipation flows more than double during the pandemic and also show sizable gender gaps pointing to a greater rise for women with children.

We conclude by discussing some of the continuing impacts of the pandemic on the labor market. In particular, we focus on what the elements of family status, occupation, and gender might foretell about whether the US economy is likely to experience another "jobless recovery," and how the newly established patterns of remote work may affect gender wage gaps looking forward.

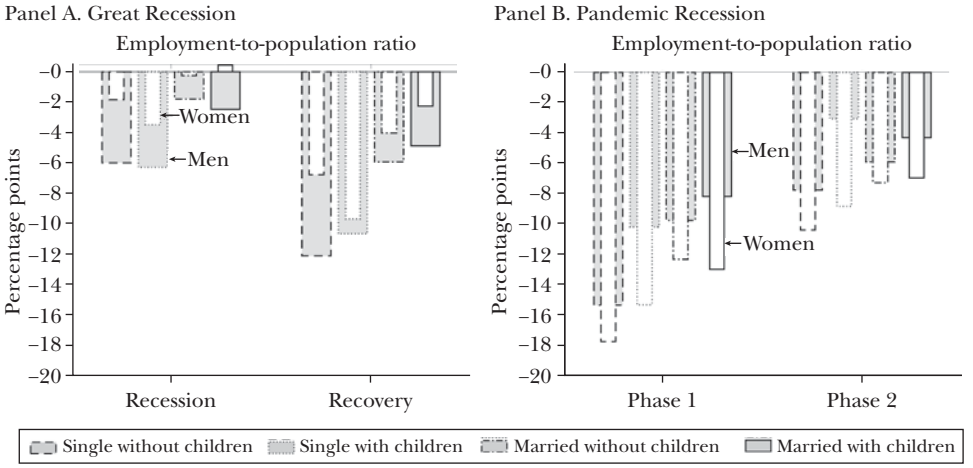
Employment by Gender and Family Status

On the supply side of the labor market, the lower cyclicality of female employment applies to married individuals during past recessions. It is related to household insurance via labor supply, sometimes also known as the "added-worker" effect. The premise of this mechanism is that when one partner is at risk of earnings loss or unemployment, for example during a recession or because of a plant closing, the other partner increases their labor supply. Since the first study isolating this mechanism, Lundberg (1985), a variety of contributions have confirmed the importance of this channel. Those looking for more recent starting points in this literature might begin with Shore (2010), who examines risk sharing within marriage over the business cycle and finds that incomes of husbands and wives are less positively correlated in recessions. Additionally, in cohorts of couples who had been married through relatively bad times, high-earning husbands tend to be married to low-earning wives and vice versa, with a large and statistically significant effect. Blundell, Pistaferri, and Saporta-Eksten (2016) examine the link between wage and consumption inequality using a structurally estimated life-cycle model incorporating consumption and family labor supply decisions, and find a sizable role for household insurance via wives' labor supply.

At the macroeconomic level, Albanesi (2019) shows that this channel renders women's labor supply countercyclical. Ellieroth (2019) finds that married women are less likely to leave the labor force in recessions. She shows that this form of

Figure 2

Change in the Employment-to-Population Ratio Relative to Pre-Recession during the Great Recession and the Pandemic Recession by Gender and Family Status



Source: Author's calculations from CPS.

Note: For the Great Recession, pre-recession corresponds to March–November 2007, Recession to December 2007–June 2009, and Recovery to July 2009–July 2012. For the pandemic recession, pre-recession corresponds to February 2020, Phase 1 to March–May 2020 and Phase 2 to June–November 2020.

precautionary labor supply in response to the higher threat of job loss experienced by their husbands accounts for 30 percent of women's low cyclicality of employment.[1]

Employment Declines in the Pandemic and the Great Recession

To illustrate how the employment losses of men and women during the pandemic recession differed from earlier recessions, we compare the pandemic recession to the Great Recession, which had a typical pattern. Figure 2 shows the change in the employment-to-population ratio by gender and family status during the pandemic recession in 2020 and the Great Recession. For each period, two changes are shown. For the Great Recession, the first change is from the pre-recession phase corresponding to the period between March 2007 and November 2007 to the recession phase from December 2007 to June 2009 (which corresponds to the dates determined by the Business Cycle Dating Committee of the National Bureau of Economic Research). The recovery phase is the change from the original pre-recession period all the way to the following recovery from July 2009 to July 2012. For the pandemic recession, both changes are relative to February 2020. We consider two time periods: Phase 1, comprising March, April,

[1] Other recent papers develop quantitative models capturing the implications of marital risk-sharing for consumption smoothing and welfare. See Albanesi (2019) for a review of this literature.

and May 2020, when the pandemic started and the strictest mitigation measures were in place, and Phase 2, from June to November 2020, with less stringent mitigation measures. Each of these changes are broken down into four categories by family status: single without children, single with children, married without children, and married with children. Then within each of these family categories, the inner portion of the bar shows the change for women, while the outer portion shows the change for men.

During the Great Recession (as in the previous recessions before that), the decline in women's employment is sizably smaller than men's for every family group. In the recession phase, among single workers without children, employment falls by 6 percentage points for men and only by 2 percentage points for women. Among single with children, the decline is 6.1 percentage points for men and 2.7 percentage points for women. For married men without children, the decline is 2 percentage points whereas women in this category employment is virtually unchanged. For married men with children, the employment-to-population ratio declines by 2.4 percentage points, while it rises by 0.2 percentage points for women. In looking at the change from the pre-recession period to the recovery phase, both women and men experience larger declines in employment, but the decline for women is one-half to one-third smaller in magnitude compared to men in each demographic group.[2]

During the pandemic recession, the decline in employment is larger for women than for men in every family group in both comparisons. In the Phase 1 comparison, single men without children employment declines by approximately 15 percentage points, whereas the decline is 18 percentage points for comparable women. For single men with children, the decline is 10 percentage points, while it is 15.5 percentage points for single women with children. For married men without children, the decline is 10 percentage points, but it is equal to 12.5 percentage points for married women without children. Finally, for married men with children, the decline is 8.5 percentage points, while for comparable women the employment-to-population ratio declined by 13 percentage points. In the Phase 2 comparison, employment continues to be well below pre-pandemic levels. For men, employment ranges between 8 and 3 percentage points below pre-pandemic levels depending on family status, and for women between 11 and 8 percentage points lower, with the largest gender gaps among workers with children. Among single workers with children, the employment decline for women relative to pre-pandemic levels is more than twice as large as for men, while for married workers with children it is approximately 50 percent larger.

[2] In the online Appendix available with this paper at the *JEP* website, we use yearly data on prime-age workers from the Current Population Survey to capture the variation in the employment-to-population ratio associated with cyclical variations in GDP in 1976–2019. We confirm the standard finding of lower cyclicality for women's employment. We also examine the cyclicality of men and women by marital status and presence of children, and confirm the patterns of recessions discussed in the text and in the context of the Great Recession.

Discussion

During 2020, women—especially those with children—experienced a substantial reduction in employment compared to men, contrary to the pattern that prevailed in previous recessions. Both labor demand and supply factors likely contributed to this behavior. Women are more likely to be employed in service-providing industries and service occupations. These tend to be less cyclical compared to goods-producing industries and production occupations that employ a larger share of men, and Albanesi and Şahin (2018) show that this accounts for most of the difference in the loss of employment during recessions since 1990. The occupation and industry distribution by gender does not vary by marital status (Cortes and Pan 2018), and thus can help explain why both for single and married workers employment is less cyclical for women. However, during the COVID-19, infection risk was most severe in the service sector, leading to a large reduction in demand for services, due to government imposed mitigation measure and customer response to infection risk. The overrepresentation of women in service jobs likely accounts for a sizable fraction of their decline in employment relative to men.

Another unique factor associated with the pandemic recession was the increased childcare needs associated with the disruptions to school activities, which may have contributed to a reduction in labor supply of parents. Why was it mothers in particular who responded to the lack of predictable in-person schooling activities in households where fathers were also present? Gender norms likely played a role. But from the perspective of an economic model of the family, this response should also be driven by differences in the opportunity cost as measured by wages. In the United States and other advanced economies, there is a substantial "child penalty" that reduces wages for women when, and even before, they become mothers and throughout the course of their lifetime. The penalty is driven by a combination of occupational choices, labor supply on the extensive and intensive margin, that begin well before women have children (Kleven, Landais, and Søgaard 2019; Adda, Dustmann, and Stevens 2017). The mean child penalty can be decomposed into explained effects, such as differences in mean values of background characteristics like education and race, and unexplained effects, which include the child penalty and different returns on non-child characteristics for mothers, compared to non-mothers or men. In a recent sample of such work, Cortes and Pan (2020) estimate that the long-run child penalty—three years or more after having the first child—for US mothers is 39 percent, and they also find that child-related penalties account for two-thirds of the overall gender wage gap in the last decade.

Given the child penalty, most working mothers at the start of the pandemic were likely to be earning less than their partners, and for those couples the optimal response to the increased child supervision needs was for mothers to reduce labor supply.[3] In addition, Cajner et al. (2020) show that employment losses were

[3] It is hard to test the implications of this hypothesis, given the unavailability of real time data on earnings by labor market transitions at high frequency. Our own preliminary work looking at monthly data on earnings from the Current Population Survey earners' study suggests that the wife/husband earning

concentrated disproportionately among lower wage workers at the beginning of the pandemic, and Chetty et al. (2020) find that by the fall of 2020, lower wage workers' employment was still more than 20 percent below pre-pandemic values, with a much larger recovery for higher wage workers. Given that the child penalty tends to relegate women to jobs and occupations at the lower end of the wage distribution, it may have also played a role in their disproportionate loss of employment. The next section considers gender differences in occupations during the pandemic recession.

Exposure to the COVID-19 Recession by Occupation

To determine exposure to the COVID-19 recession, we classify occupations along two dimensions based on their flexibility and contact intensity. The distinction between flexible and inflexible occupations is made according to whether the occupation can be carried out remotely: flexible occupations include occupations that allow their employees to work remotely, whereas inflexible occupations involve outdoor activities or require operating on site equipment. The distinction between high-contact and low-contact occupations is based on workers' physical proximity to customers or coworkers while on the job. We then document the distribution by gender across these groups of occupations.

We then measure flexibility and contact intensity using data from the Occupational Information Network (O*NET). The O*NET survey started in 1998 and we use the most recent version published in February 2020. O*NET asks a random sample of US workers in each occupation various questions about typical work activities required in their occupations. To measure occupations' flexibility, we consider 15 questions designed to elicit whether workers are performing tasks that can be executed remotely, or whether they are bound to their work location by the need to operate or inspect equipment. Respondents answer each question on an ordinal scale of one to five, and we take the average across respondents' answers. To compute the contact intensity measure, we use a question asking about physical proximity to other people while working. Again, respondents answer on a scale of one to five, described as follows: 1) beyond 100 ft., 2) private office, 3) shared office, 4) at arm's length, 5) near touching.[4]

---

ratio for two-earner couples increased from 84 percent in Jan–Feb 2020 to 88 percent in summer 2020, which was 5 percentage points above the 83 percent level in summer 2019. In addition, the average weekly earnings for married men whose wives are not in the labor force increased by 11 percent in summer 2020 compared to a year prior, while it only increased by 7 percent for married men whose wives are employed, which could have resulted from the fact that wives of higher-earning men left the labor market during the pandemic. This evidence is consistent with the hypothesis that mothers' lower wages can explain their departure from the workforce in response to the child care demands driven by the pandemic, but a deeper analysis is warranted as more data becomes available.

[4]Table 6 in the online Appendix presents the measures of flexibility and contact intensity by occupations' major groups.

Table 1
Occupation Classification

	Flexible	*Inflexible*
High-contact	Education, Training, and Library	Healthcare Practitioners and Technical Healthcare Support Food Preparation and Serving Personal Care and Service
Low-contact	Management Business Computer and Mathematical Architecture and Engineering Life, Physical, and Social Science Community and Social Services Legal Arts, Design, Entertainment, Sports, and Media Sales and Related Office and Administrative	Protective Service Building and Grounds Cleaning and Maintenance Farming, Fishing, and Forestry Construction Trades, Extraction Installation, Maintenance, and Repair Production Transportation and Material Moving

Source: Author's calculations based on O*NET.
Note: Occupations are inflexible if their inflexibility score is above the median and flexible otherwise. Occupations are high-contact if the contact intensity score corresponds to a distance of less than 6 feet. Flexibility scores and contact intensity scores are reported in Table 6 in the online Appendix.

We classify occupations as inflexible if their inflexibility score is above the median and as flexible otherwise. Similarly, we consider the occupation to be high-contact occupation if an average respondent says that they work at arm's length or closer to other people, which is closer than safe social distancing distance of 6 feet. Based on these two criteria, we aggregate occupation groups into four categories: flexible and high-contact, flexible and low-contact, inflexible and high-contact, or inflexible and low-contact. This grouping is reported in Table 1.[5] Flexible/high-contact occupation comprise mainly education jobs, while flexible/low-contact occupations comprise managerial and professional occupations. Inflexible/high-contact occupations are dominated by healthcare and services, both personal and hospitality. Finally, the inflexible/low-contact category comprises most production, protection, and transportation occupations, as well as construction and farming.

Table 2 reports the distribution of workers by gender across occupations for four categories defined in Table 1. The inflexible/high-contact occupations are the most vulnerable to the COVID-19 shock and are dominated by female workers. We find that 26 percent of female workers are employed in occupations that are inflexible/high-contact, while only 6 percent of men work in these occupations, corresponding to a female share of employment in these occupations of 73 percent. Flexible/high-contact occupations also exhibit a high female share at 76 percent. Male workers

[5]Dingel and Neiman (2020) classify occupations based on the ability of working remotely. Mongey, Pilossoph, and Weinberg (2020) also consider how workers in different occupations are affected by social distancing policies.

Table 2
Occupational Distribution by Gender

Group	*Employed women*	*Employed men*	*Total employed*	*Female share*
Flexible, High-contact	10	3	6	76
Flexible, Low-contact	53	48	51	50
Inflexible, High-contact	26	9	17	73
Inflexible, Low-contact	11	40	26	19

Source: Author's calculations based on February 2020 CPS.
Note: All values in percentage.

are disproportionately represented in inflexible, low-contact occupations, with 40 percent of male workers but only 11 percent of female workers employed in these occupations, with a female share of employment of only 19 percent in this category. Flexible/low-contact occupations are the largest category, accounting for 51 percent of overall employment, specifically 53 percent of female employment and 48 percent of male employment, with a female share of 50 percent.

We calculated the variation in employment-to-population ratio for these four sets of occupations starting in February 2020 and comparing each month in 2020 to the corresponding month in the previous year, which should help to account for any seasonality in employment variation by occupation.[6] Figure 3 displays the results in the aggregate and by gender for each group of occupations. Inflexible/high-contact occupations show the largest overall decline in employment, reaching a trough of –38 percent in April, and only recovering to –12 percent by September 2020, with further declines by the end of the year. Men's employment fell by 8 percentage points more in April compared to women. Though it recovered to approximately 18 percentage lower relative to one year prior by July, it stayed at that level or lower through the fall. By contrast for women, employment was only 13 percent lower relative to one year prior by August 2020, and it remained mostly stable through fall 2020.

Inflexible/low-contact occupations are the second worst hit, with an overall decline in employment of close to 30 percent in April 2020, though employment for these occupations is only 5 percent lower than one year prior by the fall. For these occupations, women's employment dropped to 42 percent relative to one year prior in April, much larger than the 25 percent fall for men. Men's employment recovered slowly but steadily, reaching a level 10 percent lower than one year prior by September 2020. Women's employment also eventually recovered to that level, though there was a period of further reduction in early fall. Inflexible/high-contact and inflexible/low-contact occupations comprise most workers deemed essential, even if this designation varies by state (Blau, Koebe, and Meyerhofer 2020). Yet, these two categories experience the biggest decline in employment.

[6]Of the four categories here, flexible/high-contact occupations—which are the main location for teaching-related occupations—are the only category displaying seasonal variation.

Employment in flexible/high-contact occupations dropped to a low of –17 percent relative to one year prior in April 2020 but recovered rapidly, and has remained 2–8 percent lower than one year prior in the summer and fall. Employment for women fell by approximately 5 percentage points more in April and in the fall compared to men, though the drop in employment for women was smaller than for men in the summer.

Finally, flexible/low-contact occupations, which account for the biggest share of employment, were the least impacted, with a drop in employment of –9 percent relative to one year prior in April, and a recovery to 2–4 percent lower relative to one year prior from June onward. The drop in employment in the spring was similar for men and women, though female employment remained approximately 1 percentage point lower in the summer and fall 2020 compared to men.

Two patterns emerge from these results. First, for flexible/low-contact occupations, the recovery in employment was smaller for women. Though the percent difference by gender is small, it is still notable as this category accounts for the largest share of female employment and therefore affects a large segment of the female workforce. Second, for inflexible occupations, workers with the lowest representation by gender lost more jobs. This pattern may arise due to negative selection of male workers into female-dominated inflexible/high-contact occupations and of female workers into the male-dominated inflexible/low-contact occupations. Additionally, essential frontline workers are concentrated in inflexible/low-contact occupations and because, as documented in Blau, Koebe, and Meyerhofer (2020), they are more likely to be men—this may contribute to the greater decline of employment for women in this category.

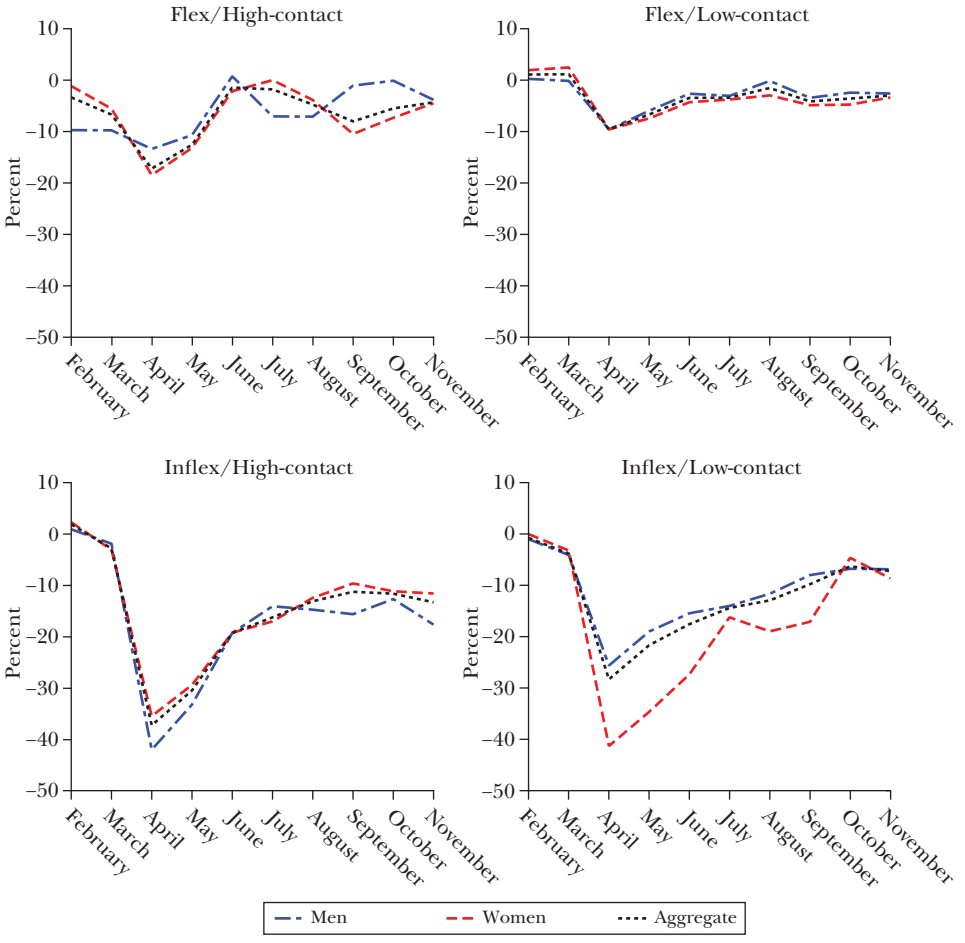
In the next section, we consider interactions between occupation and family status. The online Appendix provides more details, documenting the cyclical behavior of employment for the occupational categories we define by gender and marital status. Flexible/low-contact occupations are the least cyclical for all workers, followed by inflexible/high-contact occupations, whereas inflexible/low-contact occupations display the highest cyclicality. Albanesi et al. (2020) show that this variation is driven by differences in the skill distribution, with workers without a college degree disproportionately represented in inflexible/low-contact occupations. Additionally, workers in high-contact occupations tend to be employed in service-providing industries, which are less cyclical than goods-producing industries. However, large differences remain within occupations by gender and marital status, reflecting the aggregate pattern documented for the Great Recession.

Comparing Demand and Supply Effects

The behavior of employment over the course of the pandemic is driven by a combination of demand and supply factors that likely differ by gender and are influenced by age and education, in addition to the presence of children. In this section,

Figure 3

Percentage Change in the Employment-to-Population Ratio from One Year Prior in 2020



Source: Authors' calculations based on CPS.

we suggest two ways of disentangling these effects. One approach looks at the data on employment by marital and parental status presented earlier and asks how much of that variation would be eliminated by adding control variables for occupation. The second approach examines flows between employment, unemployment, and participation.

The overall lesson that emerges is that while both supply-side and demand-side effects play a role in explaining the drop in employment-to-population ratio for women during the pandemic recession, supply-side factors related to marriage

Table 3
Change in Employment, Unemployment, Nonparticipation by Demographic Groups

	Employment		*Unemployment*		*Nonparticipation*	
Change since February 2020	*Phase 1*	*Phase 2*	*Phase 1*	*Phase 2*	*Phase 1*	*Phase 2*
Average without occupation controls	–5.3	–3.8	5.0	3.5	0.3	0.3
Share women	64.3	62.2	63.7	61.1	72.7	78.0
Average with occupation controls	–3.6	–3.1	3.5	3.0	0.1	0.1
Share women	65.9	61.6	65.8	60.0	69.5	121.6

Source: Authors' calculations based on CPS.
Note: The table reports selected estimates from the equation in the text for employment, unemployment, and nonparticipation. The full set of estimates are reported in the online Appendix. Phases of the pandemic correspond to March to May for Phase 1, June to November for Phase 2. The average effect is obtained by summing the contribution of each demographic group, obtained by multiplying the corresponding estimated effect for each phase of the pandemic with the group's population share in February 2020. The average effect is reported for the specification without and with occupational controls. In each case, "Share women" is the sum of all female contributions divided by the average effect for the specification with occupation controls. Population shares In February 2020 are reported in the online Appendix. All values in percentage.

and children are associated with roughly about two-thirds of the shift, while occupational changes are associated with the other one-third.

Regression Framework

To examine the dynamics of employment over the course of 2020, we estimate the following regression:

$$Y_{i,t} = \alpha + \sum_{\tau=1}^{2} \beta_\tau \times I(\tau) + \gamma I^i(f) + \delta I^i(m) + \eta I^i(c) + \nu X_t^i + \epsilon_{i,t}.$$

where i indexes an individual and τ is an indicator variable for one of two phases of the pandemic, which are the same as defined earlier for Figure 1, with $\tau = 1$ corresponding to March to May, $\tau = 2$ corresponding to June to November. The variable $I^i(f)$ is a dummy for gender, equal to 1 for female, $I^i(m)$ is a dummy for marital status, equal to 1 for married, and $I^i(c)$ is a dummy for children under the age of 12 present, equal to 1 if they are, and X_t^i include a set of controls for age, educational attainment and, in some specifications, occupation, as categorized in the previous section. Additionally, we include a full set of interactions between the phase effects and the gender, marital status, and presence of children dummies, and the age, education and occupation controls.

With this approach, the coefficients β_τ estimate the effect of each phase of the pandemic on the dependent variable. The coefficients on the interactions estimate the differential impact of the pandemic on individuals by gender, marital status and presence of children in each phase of the pandemic. The estimated value of α will

be the average value of the dependent variable for male, single, childless individuals in February 2020. A full description of the regression framework, data and results is reported in the online Appendix available with this paper at the *JEP* website.

We focus here on differences between female and male employment in each of the phases of the pandemic, relative to February 2020, and calculated from the change for each demographic group weighted by the corresponding population shares for February 2020. In Table 3, we report these estimates for the specification with and without occupation controls. We also report the share of the change accounted for by women for each specification.

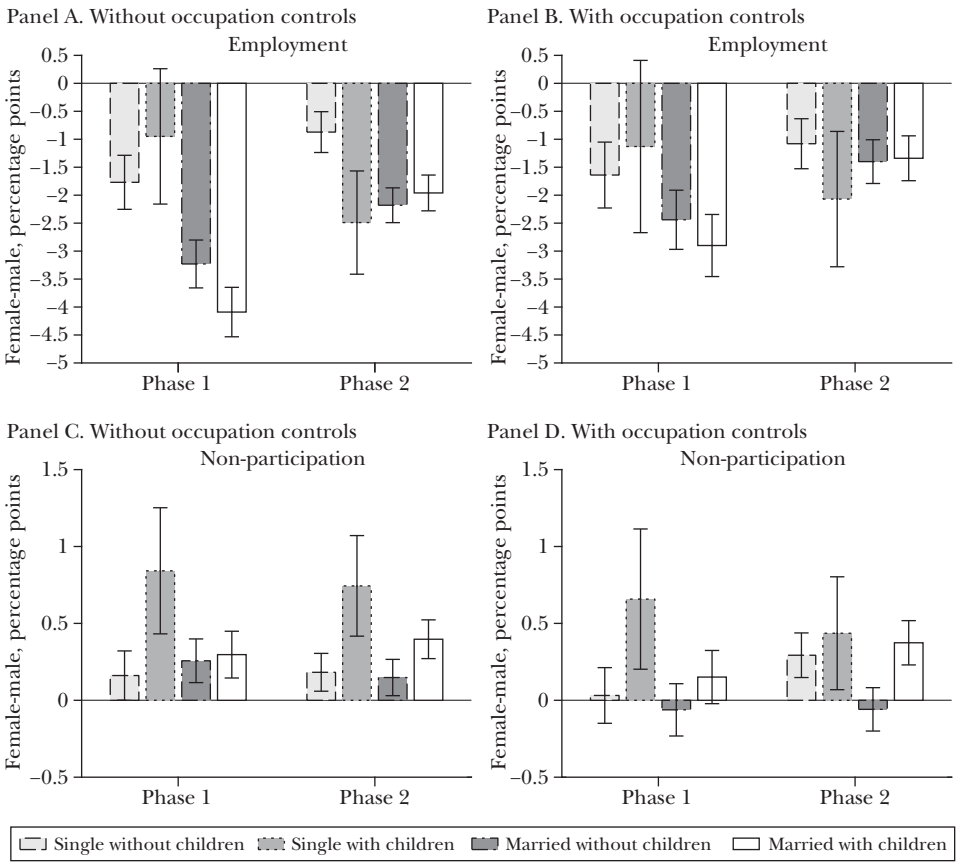
Without occupation controls, employment declined on average by 5.3 percent in Phase 1 and 3.8 percent in Phase 2; controlling for occupations, the declines were smaller at 3.6 percent and 3.1 percent. This suggests that the occupational distribution can account for over one-third of the decline in employment in Phase 1 and approximately one-fifth in Phase 2. The share of this change accounted for by women is similar with and without occupational controls, ranging from 62 to 66 percent, much larger than women's share in the population in February 2020, which was 52 percent.

The two panels in Figure 4 display the gender differences in the changes in employment in the two phases of the pandemic by demographic group, with and without occupation controls. (The values of the bars in the unadjusted figure will not match those from Figure 2 in the earlier discussion, because those values are not adjusted for age or education.) The figure shows that women in all demographic groups suffer larger losses in employment compared to men at every stage of the pandemic, with the biggest gender differences estimated for married women with children, whose employment falls by an additional 4 percentage points compared to men in that category in Phase 1. For married women without children, employment falls by 3 percentage points more than for men in that category over the same time period. Among single parents, women's employment falls by 1 percentage point more than men's in Phase 1 and by 2.5 percentage points more in Phase 2, and among single individuals without children, women's employment losses are 2 percentage points larger than men's in Phase 1 and 1 percentage point larger in Phase 2.

Controlling for occupations attenuates the gender differences in employment losses by about one-quarter to one-third in both phases of the pandemic. These estimates suggest that the occupation distribution plays a limited role in accounting for the gender gaps in the drop in employment.

We calculated similar regressions looking at patterns of unemployment and labor force nonparticipation. We find a very similar pattern for unemployment, both in terms of the average response, the contribution of occupation controls and for the gender wage gaps by demographic group. For non-participation, without controlling for occupation, gender differences are sizable for parents, particularly single parents, and more pronounced in Phase 2. Controlling for occupation, the gender gap in the rise in participation is 0.5 percentage points for parents, which is attenuated by about one-third for single parents relative to no occupation controls,

Figure 4
Female–Male Difference in Changes in Employment since February 2020, Estimated with and without Occupation Controls



Source: Author's calculations from Current Popoulation Survey data, using equation in text.
Note: See note to Table 3. Error bars denote 90 percent confidence intervals.

but not for married parents in Phase 2. These results suggest that the occupation distribution plays little role in accounting for the rise in nonparticipation for mothers with children, particularly married mothers, relative to men.[7]

[7]Figure 8 in the online Appendix, available with this paper at the *JEP* website, displays the gender differences in the change in unemployment by demographic group with and without occupation controls. We also report a set of parallel regressions for employment, unemployment, and non-participation using data from the Great Recession. The results suggest that the occupational distribution is a significant factor in women's smaller losses in employment compared to men in this period.

Evidence from Gross Labor Flows

To explore the potential role of labor demand and supply factors during the pandemic, we also examine gross labor market flows between employment, unemployment, and labor force participation. To capture the impact of demand factors, we consider the employment-to-unemployment flow and the unemployment-to-employment flow. The employment-to-unemployment flow is commonly interpreted as a measure of job destruction and usually rises dramatically at the start of recessions. The reverse measures the rate at which the unemployed find jobs, and it tends to fall dramatically in recessions and rise during recoveries. Because the unemployed are willing to but can't find work, the flows in and out of unemployment are more associated with the number of jobs available in the labor market rather than individual workers' decisions to supply labor. In contrast, the flows into nonparticipation reflect workers' voluntary choices to leave labor market.

To capture the impact of labor supply factors, we consider the employment-to-nonparticipation flow and the unemployment-to-nonparticipation flow. The first captures voluntarily quits, while the second is often interpreted as a key measure of labor market attachment during recessions. Krusell et al. (2017) provide detailed documentation of the cyclical properties of gross job flows in the United States.[8]

The estimates of the effect of the pandemic on these flows by demographic group are reported in Table 4. Overall, we find that employment-to-unemployment flows rise by 2.9 percentage points in Phase 1 and 1.2 percentage points in Phase 2. Controlling for occupations lowers these values by one-third in Phase 1. These are large changes, as on average monthly employment-to-unemployment flows range between 1.5 and 2 percentage points for men and 1 and 1.5 percentage points for women in 1976–2007 (Albanesi and Şahin 2018).

Women contribute to 65 percent of this rise in Phase 1 and 67 percent in Phase 2, and the female share declines only modestly in Phase 2 with occupation controls, suggesting that the occupation distribution plays a small role in accounting for gender gaps in the change in employment-to-unemployment flows. This can be seen in Figure 5 which reports the gender gaps by family status for this variable. These gaps are substantial for all demographic groups, ranging from 1 percentage point for single without children to 2.2 percentage points for married with children in Phase 1, and from 0.5 percentage points for single without children to 1.1 percentage points for single with children in Phase 2. Controlling for occupation attenuates these gaps by at least one-third for all categories, except for single women with children, in both phases of the pandemic. These results suggest that

[8]The possibility of classification error is an important concern when analyzing gross job flows. Earlier research has found these errors to be sizable for transitions between unemployment and nonparticipation. A standard approach to correct this issue is to adjust the gross flows data using Abowd and Zellner (1985) estimates of misclassification probabilities based on resolved labor force status in reinterviews in the Current Population Survey, as in Elsby, Hobijn, and Şahin (2015). However, given the short time span of our data and the exceptional nature of the labor markets transitions during the pandemic, it is questionable that those corrections accurately capture the extent of misclassification for our sample. For that reason, we do not apply any adjustment.

Table 4
Change in Gross Labor Flows by Demographic Groups

	EU		*UE*		*EN*		*UN*	
Change since February 2020	*Phase 1*	*Phase 2*	*Phase 1*	*Phase 2*	*Phase 1*	*Phase 2*	*Phase 1*	*Phase 2*
Average without occupation controls	2.9	1.2	–0.4	–0.6	0.2	0.1	0.1	0.1
Share women	65.1	66.6	57.6	62.1	68.7	68.0	71.0	61.0
Average with occupation controls	1.8	1.1	–0.4	–0.4	0.2	0.1	0.01	0.04
Share women	66.8	58.5	72.4	86.2	55.4	85.0	120.2	76.2

Source: Authors' calculations based on CPS.
Note: The table reports selected estimates from equation in text for employment-to-unemployment, unemployment-to-employment, employment-to-non-participation and unemployment-to-non-participation. The full set of estimates are reported in the online Appendix. Phases of the pandemic correspond to March to May for Phase 1, June to November for Phase 2. The average effect is obtained by summing the contribution of each demographic group, obtained by multiplying the corresponding estimated effect for each phase of the pandemic with the group's population share in February 2020. The average effect is reported for the specification without and with occupational controls. In each case, "Share women" is the sum of all female contributions divided by the average effect for the specification with occupation controls. Population shares In February 2020 are reported in the online Appendix. All values in percentage.

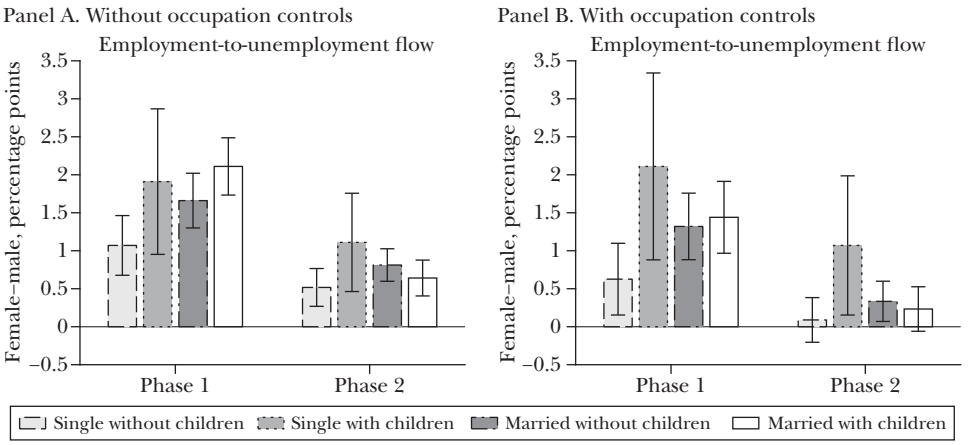
single women with children were disproportionately affected by job losses during the pandemic, beyond the effects associated with their occupation distribution.

Turning to unemployment-to-employment flows, these results show a substantial decline of 0.4 percentage points in Phase 1 and 0.6 percentage points in Phase 2, suggesting that the labor market had not yet reached a recovery phase. Occupational controls reduce the magnitude of the effect of the pandemic, though most of the effects by demographic group and corresponding gender gaps are not statistically significant.

For the flows from employment into nonparticipation, which we interpret as evidence of a supply-side shift in the labor market, we find a substantive rise during the pandemic with sizable gender differences. Employment-to-nonparticipation flows rose 0.2 percentage points in Phase 1, and by 0.1 percentage points in Phase 2, and 68 percent of this change is accounted for by women. This is a very large increase, as the average for these flows have been 0.023 for men and 0.035 for women in recent years (Albanesi and Şahin 2018). Controlling for occupation attenuates this rise only in Phase 1, and in Phase 2 increases the share of the rise accounted for by women. As shown in Figure 6, controlling for occupations, the gender differences in the change in the employment-to-nonparticipation flows are mostly driven by single women with children, for whom the rise is percentage points higher than comparable men in Phase 1 and 0.6 percentage points higher in Phase 2. Married women with children also experience a larger increase in this flow compared to men in the same demographic group in Phase 2. Turning to unemployment-to-nonparticipation flows, we find that there is an average increase of 0.1 percentage points during the pandemic. This rise is disproportionately accounted for by women, controlling for occupation,

Figure 5

Female–Male Difference in Changes in EU Flows since February 2020, Estimated with and without Occupation Controls



Source: Author's calculations from Current Population Survey data, using equation in text.
Notes: See note to Table 4. Error bars denote 90 percent confidence intervals.

with a large and significant gender gap among married workers with children in Phase 2.

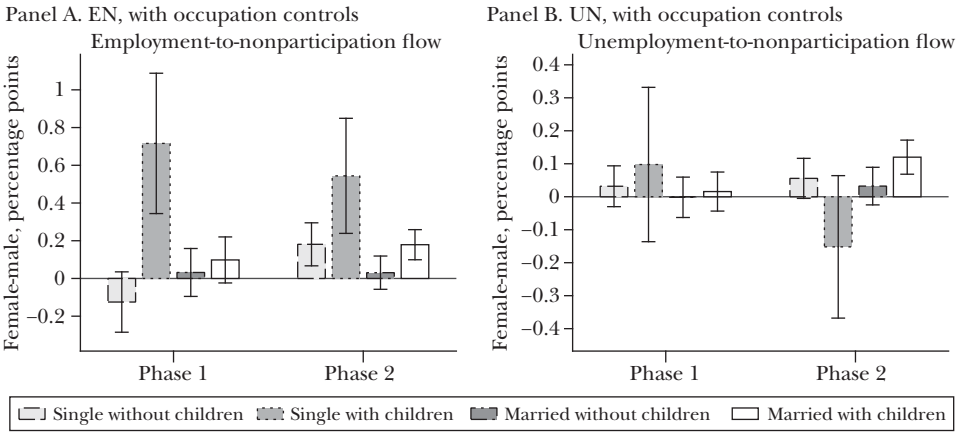
The disproportionate rise in flows into nonparticipation for women during the pandemic is striking, as it follows several decades of continued convergence in these flows across genders (Albanesi and Şahin 2018). Historically, women have exhibited higher employment-to-non-participation flows than men, with most of the difference accounted for by men's higher rate of job-to-job transition, with the gap mostly accounted for by women's tendency to exit the labor force temporarily after the birth of a child (Royalty 1998). However, as women's participation has grown, there has been a decline in their employment-to-nonparticipation transition rates. Additionally, as shown in Ellieroth (2019), these flows tend to fall for married women in recessions. Controlling for occupation, we find that unemployment-to-nonparticipation flows increase more for women, though the only significant gender gap is for married workers with children. It bears noting that historically these flows have been higher for women than for men. However, women's increased labor force attachment (Albanesi and Şahin 2018) has considerably contributed to increasing the average duration of unemployment in the United States since the early 1990s (Abraham and Shimer 2001).

Continuing Impacts

As we look forward to the end of the pandemic, one critical question is whether employment will return to pre-pandemic levels and jobs that were lost during the

Figure 6

Female–Male Difference in Changes in Employment-to-Nonparticipation and Unemployment-to-Nonparticipation Flows since February 2020, Estimated with Occupation Controls



Source: Author's calculations from Current Population Survey data, using equation in the text. .
Note: See note to Table 4. Error bars denote 90 percent confidence intervals.

pandemic will be reinstated. Since the 1990–1991 recession, the US economy has experienced "jobless recoveries"—that is, even as GDP and aggregate demand rebound from the trough of the cycle, labor market indicators continue to stagnate and employment struggles to attain pre-recession levels. After the 1990–1991 recession ended in March 1991, for example, it took until February 1993 for employment to reach its pre-recession peak. After the 2001 recession ended in November, employment only reached its pre-recession peak in October 2003. And after the Great Recession in June 2009, it took until May 2014 for total employment to reach its pre-recession peak.

Two main factors appear to be behind this phenomenon. First, Albanesi (2019) argues that the subdued behavior of employment during recoveries since the 1990s is driven by the flattening of female labor force participation. Recoveries before the 1990s have commonly been jobless for men, but as long as female labor force participation was rising briskly, female employment tended to grow very rapidly in recoveries. But as the rise in female participation slowed in the 1990s, the rate of growth of women's employment during recoveries has been similar to men's in the recessions since 1990–1991. If the recovery from the pandemic is associated with a rebound of female participation to pre-pandemic levels, the rebound in aggregate employment may be faster compared to recent cycles.

However, a second explanation for jobless recoveries points in the opposite direction. The hypothesis is that the slow and incomplete rebound of aggregate

employment is due to structural change leading to a long-run decline in certain areas like manufacturing employment (Groshen and Potter 2003) and routine jobs (Jaimovich and Siu 2020). The job losses associated with these slow-moving trends are concentrated in recessions, but then as the economy recovers, those jobs are not reinstated. This phenomenon affects primarily middle-skill jobs, which are particularly cyclical (Foote and Ryan 2015), and is a key mechanism through which the trend toward job polarization (Acemoglu and Autor 2011) has affected business cycles.

As we have argued, the pandemic has affected service occupations that in the past have seemed less amenable to automation. However, the pandemic has also given employers an additional incentive to embrace automation, an ongoing risk of infection that is expected to persist, as long as a substantial fraction of the world population remain susceptible to the coronavirus. Machines and software will not fall ill. Are jobs that were lost during the pandemic recession more or less susceptible to automation?

One way to measure the susceptibility to automation by occupation is Routine Task-Intensity (RTI), an index developed in Autor and Dorn (2013) that calculates the routine, manual, and abstract task inputs in each occupation based on job task requirements from the Dictionary of Occupational Titles (DOT). Higher values of RTI correspond to higher susceptibility to automation. Earlier in this paper we focused on four main categories of occupations. We looked at the share of occupations in each group with above median RTI and the share of pre-pandemic employment accounted for by these occupations, and show some results in Table 5.[9]

For inflexible/low-contact occupations, the most exposed to standard recessions, 22 percent of workers are employed in high-RTI jobs. For the inflexible/high-contact occupations, the category most affected by the pandemic, 34 percent of workers are employed in high-RTI positions. The most automatable occupational category with 49 percent of employed in high-RTI jobs is the flexible/low-contact, as it includes Office and Administrative and Sales and Related occupations, which are cognitive and routine. The least automatable group of occupations is flexible/high-contact, comprised of Education, Training and Library occupations. Only 0.2 percent of workers are in highly automatable jobs in this category. These findings suggest that even healthcare and personal service occupations are susceptible to automation, leaving open the possibility that employment losses in those occupations may not be fully reversed as the broader economy recovers from the pandemic.

Finally, women's employment losses from the pandemic may have longer-term effects. In the past, mothers who leave the labor force temporarily to take care of children have experienced substantial losses to wages and lifetime earnings. Adda,

[9]For details of the calculations presented here, see the online Appendix. RTI is defined as log(routine task input)- log(abstract task input)-log(manual task input). Some occupations do not have an RTI score. For the categories used in Table 5 in the text, the fraction of workers without an RTI score is 2 percent for flexible/low-contact, 8 percent for inflexible/high-contact occupations, and 6 percent for inflexible/low-contact occupations

Table 5
Susceptibility to Automation by Occupation

Occupation	*Percent employed in High-RTI*
Flexible, High-contact	0.2
Flexible, Low-contact	49.0
Inflexible, High-contact	34.3
Inflexible, Low-contact	22.0

Source: Authors' calculations based on Autor and Dorn (2013).
Note: All values in percentage.

Dustmann, and Stevens (2017) estimate that the component of the child penalty associated with "atrophy" during spells of nonparticipation, due to human capital depreciation or skill obsolescence, accounts for 13 percent of the overall gender wage gap. Additionally, employer investments in human capital and the career paths offered to women are affected by the expectation of career interruptions (Albanesi and Olivetti 2009). After many decades of increasing labor market attachment for women (Goldin 2006)—although that rise leveled out in the 1990s and has seen a small decline since then—the reduction in mothers' labor supply associated with the pandemic may reverse the slow progress made in this area.

Such effects will also interact with the extent to which remote work continues after the pandemic. Lack of flexibility has long been seen as a barrier to women's career advancement (Goldin 2014; Cortes and Pan 2019), and increased ability to work remotely, which is expected to continue after the pandemic when child care needs are normalized, may benefit women. Alon et al. (2020) conjecture that the rise in remote work may help women, as it may increase sharing of child care responsibilities with fathers now working remotely. However, the rise in mothers' nonparticipation during the pandemic suggests that, in the aggregate, this is unlikely to play a large role. In addition, even as remote work has grown for most classes of workers during the pandemic, it has increased considerably more for women (Bick, Blandin, and Mertens 2020). If it is mostly women who continue to take advantage of remote work arrangements, they may be stigmatized and miss out on career advancement opportunities, particularly in highly competitive professional and managerial occupations.

■ *We are grateful to Nicholas Fleming for excellent research assistance.*

References

Abowd, John M., and Arnold Zellner. 1985. "Estimating Gross Labor-Force Flows." *Journal of Business & Economic Statistics* 3 (3): 254–83.

Abraham, Katharine G., and Robert Shimer. 2001. "Changes in Unemployment Duration and Labor Force Attachment." NBER Working paper 8513.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, Vol. 4, 1043–1171. Amsterdam: Elsevier.

Adda, Jérôome, Christian Dustmann, and Katrien Stevens. 2017. "The Career Costs of Children." *Journal of Political Economy* 125 (2): 293–337.

Albanesi, Stefania. 2019. "Changing Business Cycles: The Role of Women's Employment." NBER Working paper 25655.

Albanesi, Stefania, Rania Gihleb, Jialin Huo, and Jiyeon Kim. 2020. "Household Insurance and the Macroeconomic Impact of the Novel Coronavirus." Unpublished.

Albanesi, Stefania, and Claudia Olivetti. 2009. "Home Production, Market Production and the Gender Wage Gap: Incentives and Expectations." *Review of Economic Dynamics* 12 (1): 80–107.

Albanesi, Stefania, and Ayşegül Şahin. 2018. "The Gender Unemployment Gap." *Review of Economic Dynamics* 30: 47–67.

Alon, Titan, Matthias Doepke, Jane Olmstead-Rumsey, and Michèle Tertilt. 2020. "The Impact of COVID-19 on Gender Equality." NBER Working paper 26947.

Autor, David H., and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–97.

Bick, Alexander, Adam Blandin, and Karel Mertens. 2020. "Work from Home before and after the COVID-19 Outbreak." CEPR Discussion Paper 15000.

Blau, Francine D., Josefine Koebe, and Pamela A. Meyerhofer. 2020. "Who Are the Essential and Front-line Workers?" NBER Working Paper 27791.

Blundell, Richard, Luigi Pistaferri, and Itay Saporta-Eksten. 2016. "Consumption Inequality and Family Labor Supply." *American Economic Review* 106 (2): 387–435.

Cajner, Tomaz, Leland D. Crane, Ryan A. Decker, John Grigsby, Adrian Hamins-Puertolas, Erik Hurst, Christopher Kurz, and Ahu Yildirmaz. 2020. "The U.S. Labor Market during the Beginning of the Pandemic Recession." NBER Working paper 27159.

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Michael Stepner, and The Opportunity Insights Team. 2020. *The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data*. Cambridge, MA: NBER.

Cortes, Patricia, and Jessica Pan. 2018. "Occupation and Gender." In *The Oxford Handbook of Women and the Economy*, edited by Susan L. Averett, Laura M. Argys, and Saul D. Hoffman, 425–52. Oxford: Oxford University Press.

Cortes, Patricia, and Jessica Pan. 2019. "When Time Binds: Substitutes for Household Production, Returns to Working Long Hours, and the Skilled Gender Wage Gap." *Journal of Labor Economics* 37 (2): 351–98.

Cortes, Patricia, and Jessica Pan. 2020, October. "Children and the Remaining Gender Gaps in the Labor Market." NBER Working paper 27980.

Dingel, Jonathan I., and Brent Neiman. 2020. "How Many Jobs Can Be Done at Home?" *Journal of Public Economics* 189:1–8.

Ellieroth, Kathrin. 2019. "Spousal Insurance, Precautionary Labor Supply, and the Business Cycle." Unpublished.

Elsby, Michael W., Bart Hobijn, and Ayşegül Şahin. 2015. "On the Importance of the Participation Margin for Labor Market Fluctuations." *Journal of Monetary Economics* 72: 64–82.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles and J. Robert Warren. 2020. *Integrated Public Use Microdata Series, Current Population Survey*: Version 8.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D030.V8.0.

Foote, Christopher L, and Richard W. Ryan. 2015. "Labor Market Polarization over the Business Cycle." NBER Working paper 21030.

Goldin, Claudia. 2006. "The Quiet Revolution that Transformed Women's Employment, Education, and Family." *American Economic Review* 96 (2): 1–21.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104

(4): 1091–1119.

Groshen, Erica L., and Simon Potter. 2003. "Has Structural Change Contributed to a Jobless Recovery?" *Current Issues in Economics and Finance* 9 (8).

Jaimovich, Nir, and Henry E. Siu. 2020. "Job Polarization and Jobless Recoveries." *Review of Economics and Statistics* 102 (1): 129–47.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209.

Krusell, Per, Toshihiko Mukoyama, Richard Rogerson, and Ayşegül Şahin. 2017. "Gross Worker Flows over the Business Cycle." *American Economic Review* 107 (11): 3447–76.

Lundberg, Shelly. 1985. "The Added Worker Effect." *Journal of Labor Economics* 3 (1): 11–37.

Mongey, Simon, Laura Pilossoph, and Alex Weinberg. 2020. "Which Workers Bear the Burden of Social Distancing?" NBER Working Paper 27085.

Olsson, Jonna. 2019. "Structural Transformation of the Labor Market and the Aggregate Economy." Unpublished. AQ: Same as before, can you provide a link?

Royalty, Anne Beeson. 1998. "Job-to-Job and Job-to-Nonemployment Turnover by Gender and Education Level." *Journal of Labor Economics* 16 (2): 392–433.

Shore, Stephen H. 2010. "For better, for Worse: Intrahousehold Risk-Sharing over the Business Cycle." *Review of Economics and Statistics* 92 (3): 536–48.

US Census Bureau. 2021. 2010 Census Occupation Codes with Crosswalk. https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html.

US Department of Labor, Employment and Training Administration. 2021. O*NET 24.2 Data Dictionary. https://www.onetcenter.org/dictionary/24.2/text/.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 25–46

The Great Unequalizer: Initial Health Effects of COVID-19 in the United States

Marcella Alsan, Amitabh Chandra, and Kosali Simon

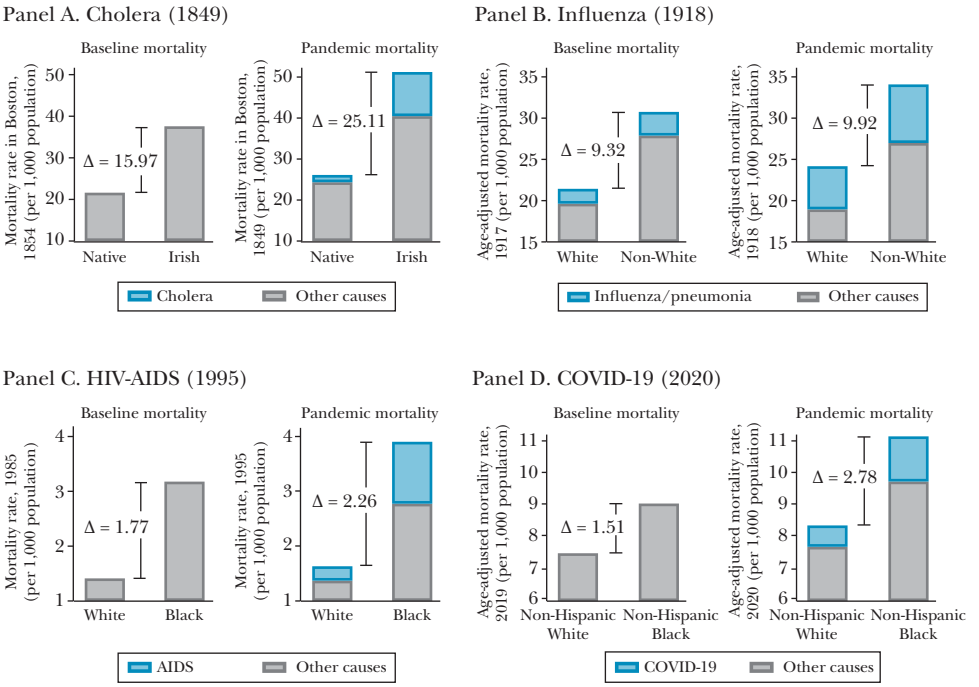
When SARS-CoV-2 first appeared, it was discussed as an equal opportunity pathogen: no one was immune, and therefore all potentially affected (Mein 2020; Krishnan, Ogunwole, and Cooper 2020). Early cases of COVID-19 disease among elites in entertainment, politics, and industry such as actor Tom Hanks, UK Prime Minister Boris Johnson, and Morgan Stanley chief executive officer James Gorman gave credence to this view.

Yet historical episodes of infectious disease are generally not experienced evenly across social strata. Rudolf Virchow (1848), the founder of modern cellular pathology and a proponent of medicine as a social science, noted that "statistics will be our standard of measurement: we will weigh life for life and see where the dead lie thicker, among the workers or the privileged." Figure 1 offers some examples of prominent novel infectious disease threats across the last two centuries in the United States, showing that the dead often indeed "lie thicker" among those less privileged. The upper-left panel shows that when cholera struck Boston in 1849, the mortality gap between native-born Bostonians and Irish immigrants and their children was

■ *Marcella Alsan is Professor of Public Policy at Harvard Kennedy School, Cambridge, Massachusetts. Amitabh Chandra is the Henry and Allison McCance Professor of Business Administration, Harvard Business School, and Ethel Zimmerman Wiener Professor of Public Policy and Director of Health Policy Research, Harvard Kennedy School of Government, both in Cambridge, Massachusetts. Kosali Simon is Herman B. Wells Endowed Professor, O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, Indiana. All three authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are marcella_alsan@hks.harvard.edu, amitabh_chandra@harvard.edu, and simonkos@iu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.25.

Figure 1
Changes in Mortality for Different Groups during Pandemics, 1849–2020



Source: For Panel A, authors' calculations from Shattuck (1846), Simonds (1850), Clark et al. (1850), Wright (1855), Curtis (1856), and Bushée (1899). For Panel B, authors' calculations from US Census Bureau (1919, 1920, 1922) and National Center for Health Statistics (2020a). For Panel C, authors' calculations from National Center for Health Statistics (2020c). For Panel D, Authors' calculations from National Center for Health Statistics (2020a, 2020e, 2021a) and US Census Bureau (2020a). See online Appendix available with this article at the *JEP* website for computational details. Also, see Appendix Figure 1 for a comparison of 2019 and 2020 mortality rates including other races and ethnicities.
Note: Figure reports mortality rate estimates by group during selected pandemics in US history and for proximate non-pandemic years. Area shaded blue denotes recorded mortality from the pandemic disease (and in the case of influenza/pneumonia, includes seasonal baseline mortality patterns from influenza and/or pneumonia), while area shaded grey denotes mortality from other causes.

about 50 percent larger than during a non-epidemic year like 1854. The upper-right panel illustrates how the age-adjusted mortality gap between White and non-White Americans rose during the influenza pandemic of 1918. The bottom-left panel displays how the HIV/AIDS pandemic worsened disparities in mortality between Black and White Americans aged 25–44. Finally, the lower-right panel shows that the age-adjusted difference in all-cause mortality rates between non-Hispanic White Americans and non-Hispanic Black Americans widened by over 80 percent during the first year of the COVID-19 pandemic.

This paper discusses the initial health effects of COVID-19 in the United States. During 2020, the first year of the pandemic, COVID-19 was recorded as the underlying or contributing cause of 378,000 deaths nationwide (Ahmad et al. 2021). The impact of COVID-19 on health, however, extends beyond its direct toll on mortality. We begin by discussing the various ways in which COVID-19's health effects have been measured as well as the role that pandemic-induced changes in the demand and supply sides of health care may have played in affecting mortality from causes other than the virus itself. Comparing the overall increase in mortality in the period 2019–2020 to the number of recorded COVID-19 deaths in 2020 indicates that the pandemic has had substantial indirect effects on health.

We next turn to examining inequality in the effect of COVID-19 on the health of different population groups. Infectious pathogens exploit both biological and social vulnerabilities, and the presentation of medical statistics can render gradients more or less conspicuous. Older age groups are particularly vulnerable to COVID-19, raising health risks for population groups with higher average ages, such as non-Hispanic Whites. However, mortality rates conditional on age are considerably higher for historically disadvantaged groups such as Black, Hispanic, and American Indians—especially due to their young average ages, these groups account for a disproportionate share of COVID-19 deaths.

The third section of the paper examines different explanations for why COVID-19 has had such unequal health effects, with a focus on racial and ethnic disparities. We provide a framework for organizing factors that contribute to the observed gradients and consider whether they are driven by preexisting differences in health risks and socioeconomic factors or by differential impacts from the same across advantaged versus disadvantaged groups. We conclude by pointing out that the patterns of health inequality seen during the pandemic mirrored those that existed in the United States prior to COVID-19 and offer thoughts about how the evolution of these gradients and resilience for the next pandemic will depend on technology, health policy, and broader social policy.

Measuring the Overall Initial Health Impact of COVID-19

The first case of SARS-CoV-2 infection in the United States was reported on January 20, 2020 (Holshue et al. 2020). From then to the end of 2020, there were an additional 20.4 million confirmed infections nationwide (Centers for Disease Control and Prevention 2021c). Of these cases, 378,000 resulted in death from COVID-19, the disease caused by SARS-CoV-2. By the measure of confirmed deaths from disease, the COVID-19 pandemic ranks among the deadliest in United States history, comparable to the 1918 influenza and HIV pandemics (Goldstein and Lee 2020).

While the health effects of the COVID-19 pandemic clearly have been significant, quantifying them is complex. A first challenge in measurement is data quality, which varies substantially by outcome of interest. Case reports of COVID-19 are

often submitted with little information on patient demographics or their severity; 37.5 percent of cases in the Centers for Disease Control and Prevention's (CDC) COVID-19 surveillance system are missing race/ethnicity data and 88.4 percent lack information on underlying health conditions (for details, see online Appendix Table 1 available with this article at the *JEP* website; Centers for Disease Control and Prevention, COVID-19 Response 2020). Although reports of "long COVID" indicate that the disease may have persistent health effects among some of those infected, systematic data on the morbidity impacts of COVID-19 are scarce (COMEBAC Study Group 2021). In the light of these data constraints, we focus on COVID-19's effects on mortality, which is a key health outcome of interest and reported by law to the CDC (National Research Council 2009).[1]

Even quantifying the impact of COVID-19 on mortality has challenges. While the number of deaths attributed to COVID-19 disease in 2020 indicates the effects have been substantial, this figure may underestimate the mortality impacts of the pandemic. An estimated 3.4 million deaths occurred in the United States during 2020, an increase of 504,000 from the 2.9 million deaths during 2019. Evolving and variable clinical presentations alongside failures in testing, both of which characterized the early pandemic, may have resulted in deaths from COVID-19 going unrecorded (Wu et al. 2020).[2] Spillover effects of COVID-19 also increased pandemic-related mortality. We describe these spillover factors below.

Effects of COVID-19 on the Demand for Health Care

Widespread avoidance of health care facilities early in the pandemic has been documented. The CDC's *Morbidity and Mortality Weekly Report* from September 2020 estimated that 41 percent of US adults had delayed or avoided medical care due to concerns about COVID-19, including 12 percent who had avoided urgent or emergency care (Czeisler et al. 2020b). In a nationally representative survey of 1,337 adults conducted in July 2020 by researchers at Johns Hopkins, 29 percent of respondents who reported needing care forwent it due to fear of viral transmission, with 7 percent forgoing care due to financial repercussions of the pandemic (Anderson et al. 2021).

Similar results are found in the Census Bureau's Household Pulse Survey, a repeated cross-section of 1.8 million US adults from the Bureau's Master Address File (US Census Bureau 2020b). Of those surveyed between April and December 2020, 37 percent reported having delayed medical care over the previous four

[1] We also do not compute period life expectancy at birth. As discussed in Goldstein and Lee (2020), "in the context of epidemic mortality, life expectancy at birth is a misleading indicator, because it implicitly assumes the epidemic is experienced each year over and over again as a person gets older." Estimates of reductions in US life expectancy during 2020 are nevertheless striking: Andrasfay and Goldman (2021) project a decline of 1.13 years in 2020 life expectancy at birth compared to a scenario without COVID-19.

[2] Symptomatic COVID-19 disease varies from mild to severe and can lead to death in a variety of ways. Pneumonia and respiratory failure are prominent final pathways, but cardiac conditions, embolic events, and systemic inflammation are also possible (Gupta et al. 2020; Malas et al. 2020; Jose and Manuel 2020; Long et al. 2020).

weeks due to the ongoing pandemic. The Pulse data indicate that delay of care followed the general contours of national COVID-19 prevalence, reaching a peak through the spring of 2020, declining in late summer, and plateauing at a lower level in early autumn before rising again. The share delaying care, however, topped 30 percent in every week the Pulse survey was fielded and stood at 35 percent in December—nine months after the national emergency began.

Effects of COVID-19 on the Supply of Health Care

On the supply side, many non-emergency interventions were suspended due to the pandemic. Shortly after President Donald Trump declared the COVID-19 outbreak a national emergency in March 2020, the Centers for Medicare and Medicaid Services (CMS) (2020) recommended the cancellation or delay of most elective surgeries and non-emergency medical, surgical, and dental procedures. National and state-level policies sought to curtail patient volume in order to conserve scarce personal protective equipment, free up beds and personnel for COVID-19 patients, and reduce SARS-CoV-2 transmission. These changes may have elevated non-COVID-19 morbidity and mortality (Chen and McGeorge 2020).

The sharp reductions in volume and increased costs providers faced during the pandemic resulted in financial distress for many health care providers. The $187 billion in federal aid allocated to providers during the crisis exhibited little relationship to COVID-19 disease burden or hospital financial health and failed to save many struggling providers even as well-resourced hospital networks, their losses cushioned with aid, engaged in a renewed wave of consolidation (Kakani et al. 2020; Abelson 2021). The closure of an estimated 8 percent of physician practices and a record number of rural hospitals, along with the higher prices and lower quality of care generally accompanying provider consolidation, may affect patient outcomes in the longer term (Physicians Foundation 2020; Basu et al. 2019; Gaynor 2018).

Overall, the pandemic's impact on the supply and demand channels described above resulted in extraordinary declines in health care utilization. Non-COVID-19 medical admissions fell by 40 percent during the first wave of COVID-19 and remained depressed nearly a year later (Birkmeyer et al. 2020; Heist, Schwartz, and Butler 2021). The implementation of policies such as stay-at-home or business closure orders may have contributed to the fall in outpatient visits (Ziedan, Simon, and Wing 2020). While substitution to telemedicine partly offset the drop of in-person care, important preventive services such as vaccinations and screenings could not be shifted online and saw precipitous declines: one study using data on over 5 million individuals with employer-sponsored insurance found decreases of 22 percent in vaccinations among children aged 0–2, 67 percent in mammograms among women aged 46–64, and 70 percent in colonoscopies among individuals aged 46–64 (Patel et al. 2021; Whaley et al. 2020). The consequences of these delays in care will likely reverberate in the form of delayed diagnosis of non-communicable disease, preventable cases of infectious disease, and strain on providers, long after the pandemic ends.

Additional Spillover Effects

The COVID-19 health crisis is also an economic crisis. Based on prior recessions, Ruhm (2000) has noted that mortality tends to be procyclical. When the opportunity cost of leisure declines, individuals have more time to exercise, prepare healthy food and, in nonpandemic times, seek medical care. The quality of health care, particularly in nursing homes, may also display cyclical fluctuations (Stevens et al. 2015). Declines in economic activity and mobility during the pandemic recession may have led to reductions in non-COVID-19 deaths, compensating in part for the rise in mortality from infectious disease and delayed care. While there is some suggestive evidence of declines in air pollution and motor vehicle deaths from the early months of the pandemic as well as a decrease in seasonal flu deaths resulting from reduced social interaction, averted deaths are likely to be relatively small in number (Cicala et al. 2020; Centers for Disease Control and Prevention 2021a).

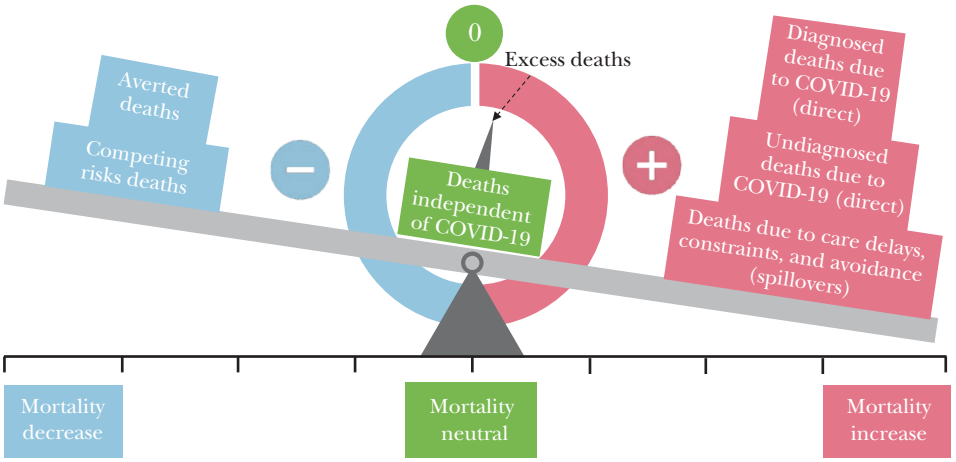
Unlike in most recessions, protective measures taken by governments and individuals to limit disease transmission during the pandemic resulted in unprecedented levels of social isolation. Disruptions in daily routines, community life, and support systems were accompanied by a troubling increase in substance use disorder: for example, the CDC's *Morbidity and Mortality Weekly Report* estimated in August 2020 that 13 percent of US adults had started or increased substance use to cope with the pandemic's effects (Czeisler et al. 2020a). Following three years of relative stability, drug overdose deaths nationwide sharply rose beginning in April 2020, the first full month of the COVID-19 national emergency, and grew through October 2020 (National Center for Health Statistics 2021b). The rise in substance abuse deaths concomitant with the pandemic suggests another avenue through which the COVID-19 crisis may have indirectly elevated mortality.

Excess Deaths: A Summary of COVID-19's Mortality Effects

The many ways through which the COVID-19 pandemic affected mortality renders precise attribution to any one cause challenging. Indeed, given that individuals are often at risk for more than one type of death, some deaths recorded as due to COVID-19 disease would have occurred even in the absence of the pandemic (Gichangi and Vach 2005). Figure 2 illustrates how the process of assessing the mortality toll of COVID-19 is complicated by the phenomena of substitution between different causes of death (competing risks), indirect deaths (spillovers), and averted deaths. The intractability of individually ascertaining the number of deaths resulting from each possible cause has motivated the use of "excess deaths" to capture the overall effect of the pandemic.

Excess deaths refer to differences between observed deaths in a particular time period and historical or expected deaths in a similar time period (National Center for Health Statistics 2020d). As Figure 1 suggests, years in which the United States experienced an infectious disease epidemic demonstrate higher death rates than contemporaneous nonepidemic years, with the increase in deaths attributable to both mortality among infected individuals and a net increase in deaths from other causes. The sum of recorded deaths from the disease and the difference in deaths

Figure 2
A Taxonomy of COVID-19's Impacts on Mortality



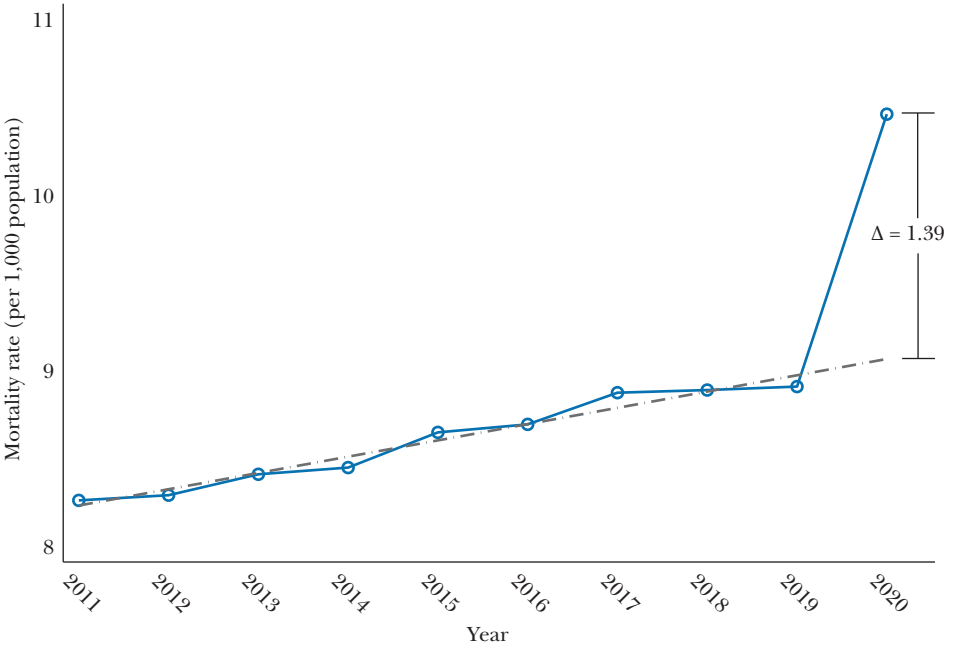
from all other causes compared to proximate time periods provide a summary statistic of the epidemic's total effect on mortality. This number of excess deaths, which can be adjusted to account for preexisting mortality trends, is typically then divided by the size of the population to yield a rate of excess mortality.

Following Polyakova et al. (2020, 2021), we estimate excess mortality as the deviation from a linear mortality trend. Figure 3 plots all-cause mortality rates among all Americans for each year from 2011 to 2020, using death data from the National Center for Health Statistics and population estimates from the American Community Survey (National Center for Health Statistics 2021a; US Census Bureau 2020a). After declining throughout much of the 20th century, mortality rates in the United States have generally risen since 2010, in part due to the nation's aging population. The number of deaths recorded in 2020, however, was far above the number expected based on prior trends. The deviation in the 2020 all-cause mortality rate from the 2011–2019 trend was 1.39 deaths per 1,000 population, or a 15.4 percent increase relative to trend. By comparison, the mortality rate from COVID-19 disease in 2020 was 1.08 per 1,000 population, suggesting that spillovers contributed to elevated mortality during the pandemic.

Measuring COVID-19 Health Inequality

The health consequences of the COVID-19 crisis affected Americans of all backgrounds, with over half of respondents in a Pew Research Center survey reporting that they knew someone personally who had died or been hospitalized due to the disease (Funk and Tyson 2020). The health toll of the pandemic, however, fell most

Figure 3
All-Cause Mortality Rates in the United States, 2011–2020



Source: Authors' calculations from National Center for Health Statistics (2020a, 2020e, 2021a) and US Census Bureau (2020a).
Note: Figure plots mortality rates from all causes for the United States from 2011 through 2020. The difference in the 2020 mortality rate compared to the 2011–2019 linear trendline is labeled. Mortality rates are not adjusted for age.

heavily on Black, American Indian, and Hispanic individuals, who disproportionately bore the total mortality burden of COVID-19 in at least two ways: they died almost always at greater rates, and they died at younger ages. We examine inequality in pandemic-related mortality, with a focus on inequality by race and ethnicity, in the section below.

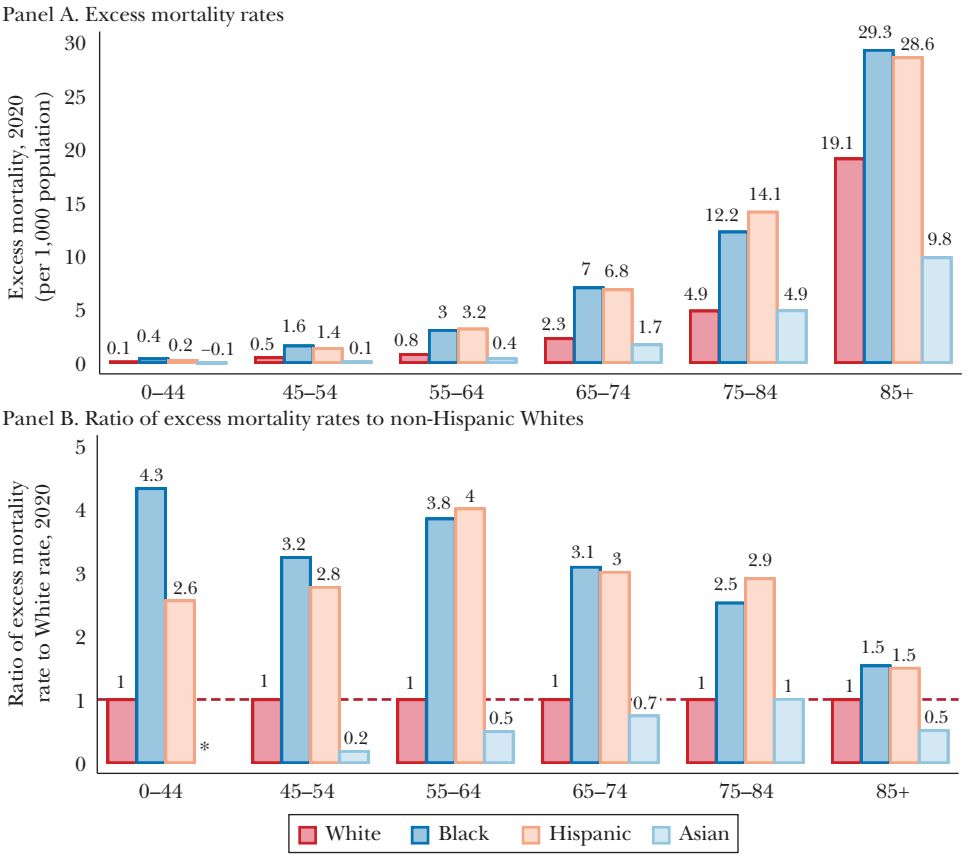
Excess Mortality by Race/Ethnicity and Age

Having estimated all-cause excess mortality during the first year of the pandemic for all Americans, we replicate this approach by race/ethnicity and age. Panel A of Figure 4 reports excess mortality rates in 2020 for non-Hispanic White, non-Hispanic Black, Hispanic, and non-Hispanic Asian Americans within six age groups (0–44, 45–54, 55–64, 65–74, 75–84, and 85 and over).[3]

[3]While American Indians and Alaska Natives (AIAN) do appear to have experienced high rates of COVID-19 infection during the pandemic, we do not assess excess mortality among these populations due to known data quality issues (Yellow Horse and Huyser 2021; National Center for Health Statistics 2021c).

Figure 4

All-Cause Excess Mortality in 2020 by Race/Ethnicity and Age Group



Source: Authors' calculations from National Center for Health Statistics (2020e, 2021a) and US Census Bureau (2020a).

Note: The rate ratio for non-Hispanic Whites is 1 by construction within each age group and is shown for reference. The rate ratio for Asians age 0–44 is omitted, because this group did not experience excess mortality.

Pandemic all-cause excess mortality rises sharply with age, largely because age is the strongest single determinant of mortality from COVID-19 (Petrilli et al. 2020) and because avoided non-COVID-19 healthcare is more likely detrimental to the oldest adults. Indeed, Americans age 70 and above experience "case fatality rates"—rates of death conditional on diagnosis—about 200 times higher than those below age 40 (for details, see online Appendix Table 2). Rates of excess mortality at

Using data from a limited set of states, a CDC *Morbidity and Mortality Weekly Report* from December 2020 found that disparities in COVID-19 death rates between AIAN and non-Hispanic White individuals were large and particularly so at younger ages (Arrazola et al. 2020).

any given age, however, vary sharply by race and ethnicity. Panel B of Figure 4 plots the ratio of non-Hispanic Black, Hispanic, and non-Hispanic Asian excess mortality rates to the non-Hispanic White excess mortality rate for each age group.

The rate ratios presented in Panel B point to two dimensions of mortality disadvantage that Black and Hispanic Americans faced. First, the Black–White and Hispanic–White ratios are above one at every age, indicating that Blacks and Hispanics experienced elevated rates of excess death compared to non-Hispanic Whites. Indeed, when all age groups are pooled and excess mortality is computed for each race and ethnicity as a whole, it is evident that Black and Hispanic populations suffered the highest rates of excess death in 2020. Panel A of Figure 5 shows that Black Americans experienced excess mortality of 2.14 deaths per 1000 population in 2020, or a 25.0 percent increase in mortality relative to trend, while Hispanic Americans in 2020 saw excess mortality of 1.44 deaths per 1000 population in 2020, or a startling 39.5 percent rise relative to trend. Non-Hispanic Whites and Asians experienced increases in mortality of 1.29 and 0.58 deaths per 1000 population respectively, which are both increases of less than 15 percent relative to trend. As Panel B of Figure indicates, these disparities widen further when excess mortality rates are age-adjusted to account for differences in age distributions between races and ethnicities — namely, the younger Hispanic and older non-Hispanic White age structures.[4]

Second, the ratios in Panel B of Figure 4 are strikingly higher at younger ages compared to older groups. The Black–White ratio in excess mortality is above four for the youngest age group of 0–44, and above three for all age groups through 65–74. Similarly, the Hispanic–White ratio in excess mortality is above 2.5 for all age groups from 0–44 to 65–74. By contrast, the Black–White and Hispanic–White excess mortality ratios for individuals age 85 and over are a comparatively low 1.5. The steep age gradient in excess mortality disparities indicates that the already high number of Black and Hispanic pandemic-related deaths disproportionately occurred among the young.

Mortality rates, whether age-adjusted or unadjusted, do not differentiate between deaths at younger or older ages. Examining death rates alone, therefore, overlooks an important dimension of inequality: differences in the number of years individuals in each group would likely have lived had they not died due to pandemic-related causes.

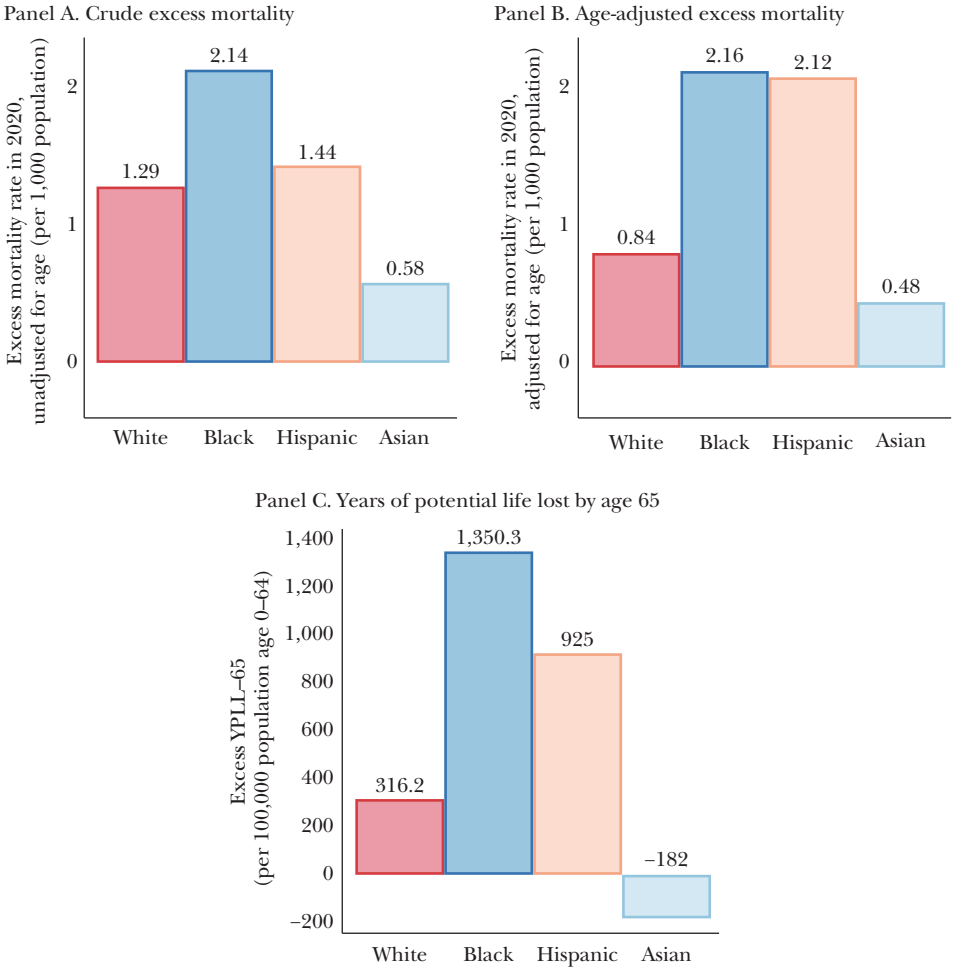
Disparities in Lost Years of Life

The concept of "years of potential life lost," or YPLL, is used in the public health literature to quantify premature mortality (Gardner and Sanborn 1990). As

[4]Age-adjusted statistics are computed by weighting deaths in different age groups among a given race or ethnicity in a manner that matches the share of each age group in the general population. Under age-adjustment, races and ethnicities with an age distribution younger than the general population have deaths at older ages weighted relatively more and deaths at younger ages weighted relatively less, whereas the converse would be true for races or ethnicities with an age distribution older than the general population

Figure 5

Measures of Racial and Ethnic Disparities in COVID-19 Pandemic Mortality



Source: Authors' calculations from National Center for Health Statistics (2020a, 2020e, 2021a) and US Census Bureau (2020a).

premature death is typically defined as a death occurring before age 65, the years of potential life lost for an individual who died prematurely is calculated by subtracting their age at death from 65, with those dying at age 65 or older assigned a years of potential life lost value of zero. Total years of potential life lost in a population is then computed by summing up the years of potential life lost among all individuals in the population who died early. This sum is usually normalized by dividing it by the number of individuals in the population under age 65. The aim of the years of potential life lost statistic is to measure life lost in years of life foregone as opposed to deaths incurred, thus providing a complementary measure to mortality rates.

As with excess death rates, we estimate "excess" years of potential life lost during the COVID-19 pandemic as the deviation in 2020 from the 2011–2019 linear trend. Panel C of Figure 5 plots excess years of potential life lost from all causes per 100,000 individuals under age 65 by race and ethnicity during 2020. Among all Americans, rates of years of potential life lost rose by 12.8 percent in 2020 compared to trend. Similarly with mortality rates, this increase was again concentrated among Black and Hispanic Americans. Black rates of years of YPLL rose in 2020 by 1,350.3 per 100,000 individuals under 65, or an increase of 19.5 percent relative to the 2011–2019 trend. Among Hispanics, YPLL rose in 2020 by 925 per 100,000 individuals under 65, or 29.2 percent relative to trend. By comparison, YPLL rates among non-Hispanic Whites increased by 316.2 per 100,000 individuals under 65, or 8.4 percent, and among Asian Americans YPLL rates fell slightly relative to trend.

Examining ratios of Black and Hispanic excess YPLL rates to the non-Hispanic White YPLL rate, in a similar manner to Figure 4, underscores the mortality disadvantage Black and Hispanic individuals have faced during the pandemic. The Black–White ratio in excess YPLL is 4.2, whereas the Hispanic–White ratio in excess years of potential life lost is 2.9. In contrast, the Black–White ratio for age-adjusted excess mortality is 2.6, and the Hispanic-White age-adjusted excess mortality ratio is 2.5. The elevated YPLL ratios suggest that not only have Black and Hispanic Americans died at greater rates during the pandemic, but those who died on average had many more years of life left to live. Far from being an equal opportunity pathogen, SARS-CoV-2 has exposed societal cleavages between less privileged and more advantaged groups.

Understanding the Unequal Health Effects of COVID-19

The COVID-19 pandemic affected some groups, particularly Black and Hispanic Americans, more than others. Why was this the case? Our goal in this section is twofold: to provide a framework for organizing the main factors that contribute to the observed disparities and to present the results of a decomposition that examines the relative importance of some of these factors.

A Framework for Understanding COVID-19 Health Inequality

We begin by focusing on deaths directly associated with COVID-19. Inequality in COVID-19 disease may be due to social determinants (such as differences in occupation, income, or education), medical determinants (including differences in comorbidities, health care quality, and insurance), and long-standing institutional features that perpetuate systemic racism and intergenerational poverty (Snowden and Graaf 2021). These factors are not exhaustive, nor are they mutually exclusive. They can, however, be mapped into an expanded model of disease transmission.

The probability of death from COVID-19 is the product of the probability of SARS-CoV-2 infection and death from COVID-19 conditional on infection. All else equal, the probability of infection rises as prevalence in the community increases

and also increases if one has more contact with others. An individual's infection probability, however, declines if more mitigating measures, such as mask-wearing, physical distancing, and vaccination, are taken.[5]

Prevalence in an individual's community and their number of contacts can be affected by social inequality, population density, and local policies. Black and Hispanic populations both live in areas with higher COVID-19 prevalence and face higher costs of reducing their number of contacts. Black and Hispanic Americans live in larger households that are more likely to be multigenerational, and are more likely to have poor housing conditions (Cohn and Passel 2018). They are also more likely to be frontline workers who must work in-person despite the risk of infection and cannot stop working or cut back on hours due to relatively low wealth levels or, particularly in the case of Hispanics, challenges in accessing federal benefits (Blau, Koebe, and Meyerhofer 2020).

Mitigation activities can help offset the risk associated with labor or leisure-related interactions. However, ability to follow public health guidance depends on access to public health information, complementary tools such as vaccines and masks, and beliefs in the credibility of health information. A survey of approximately 5,000 Americans conducted early in the pandemic showed that Black and Hispanic individuals, younger people, and men were less likely to have accurate information about COVID-19 transmission than other groups (Alsan et al. 2020a). Mitigation behavior by individuals during the pandemic has also been shaped by the dissemination of misinformation and features of the messenger, such as whether they are of the same race or ethnicity as the individual or whether they are an expert or peer (Simonov et al. 2020; Alsan et al. 2020b; Alsan and Eichmeyer 2021). Although communication with Black physicians has been shown to raise take-up of preventive health among Black Americans, just 4 percent of physicians in the United States are Black (Alsan, Garrick, and Graziani 2019).

As discussed above, mortality during the COVID-19 pandemic can be conditional on direct infection or due to indirect spillover effects. Access to quality health care is important in determining mortality, direct and indirect, from COVID-19. While higher-quality hospitals are associated with lower mortality rates, low-income Black, and Hispanic Americans obtain health care from lower-quality facilities (Jha, Orav, and Epstein 2011; Doyle, Graves, and Gruber 2019; Chandra, Kakani, and Sacarny 2020). Elevated COVID-19 caseloads in Black and Hispanic communities also contribute to non-COVID-19 excess deaths by reducing the ability of health care facilities to treat non-COVID-19 patients and causing individuals to avoid or delay necessary care due to fear of contagion. Black and Hispanic adults were also

[5] Following standard epidemiological models, the probability of SARS-CoV-2 infection can be written as $P(\text{SARS-CoV-2 infection}) = 1 - (1 - p)^{n_i(1-m_i)}$, where i refers to an individual, p represents prevalence, n is the number of contacts, and m is the proportion of mitigated contacts (Halloran 2009). Prevalence can be expressed as a function of the behavior of others around the individual ($p(n_i(1 - m_i))$), which in turn is affected by the factors noted above.

less likely at the start of the pandemic to be covered by health insurance, potentially contributing to care delays (Cohen et al. 2020).

The distribution of preexisting conditions differs by race and ethnicity, raising the probability of death conditional on COVID-19 infection and the inability to receive needed care for other chronic illnesses. Relative to non-Hispanic Whites, rates of diabetes are 1.7 and 1.8 times higher among Black and Hispanic populations (Centers for Disease Control and Prevention 2020). Rates of obesity and hypertension are similarly elevated among Black and Hispanic individuals as well (Reeves and Smith 2020). Diabetes, obesity and hypertension are conditions that increase the risk of death from COVID-19 (Centers for Disease Control and Prevention 2021b).

COVID-19 Health Inequality: A Decomposition

The potential drivers of health disparities seen during the COVID-19 pandemic are manifold. Black and Hispanic Americans are disadvantaged socioeconomically relative to non-Hispanic Whites and tend to have a greater number of comorbidities that heighten the risk of severe COVID-19 disease. Are racial and ethnic disparities in COVID-19 health outcomes driven by differences in these characteristics? Or are Black, Hispanic, and non-Hispanic White individuals differentially impacted by COVID-19 even when they possess the same attributes? We aim to examine the relative importance of each of these factors through a decomposition exercise. As datasets from the CDC largely lack detailed comorbidity data or information on individuals who have not contracted COVID-19, we obtain the necessary data from the Optum Clinformatics® Data Mart (CDM), a comprehensive commercial and Medicare Advantage claims database. In addition, we shift focus from the relatively rare outcome of mortality to COVID-19 hospitalizations.

The Optum database includes approximately 67 million unique lives of all ages across 2007–2020 and is broadly representative geographically. We include adults age 21 or older who identify as either non-Hispanic White, non-Hispanic Black, or Hispanic, who enrolled prior to July 2019, and who filed a medical claim at least once during 2019 (thus avoiding cases where comorbidities have been undiagnosed and allowing us to have three calendar quarters of data prior to the peak of the pandemic's first wave). Our analysis sample includes all enrollees who were hospitalized for COVID-19 during the first three quarters of 2020, along with a 5 percent random sample of those not hospitalized for COVID-19 as a control group.[6] Our final sample includes approximately 322,000 non-Hispanic White, 50,000 non-Hispanic Black, and 61,000 Hispanic enrollees.

We first measure whether sampled enrollees were previously diagnosed with medical conditions that increase the risk of severe illness from COVID-19: specifically, we extract information on hypertension, diabetes, obesity, cancer, heart disease, and chronic obstructive pulmonary disease based on diagnosis codes in claims filed

[6]COVID-19 testing and results are not reliably included in claims data; moreover, testing was not necessarily evenly distributed across groups (Rubin-Miller et al. 2020).

between January 1 and December 31, 2019. We also extract social and demographic information including age, sex, average educational attainment in the enrollee's census block of residence, and census division of residence. We conduct a "three-fold" Oaxaca-Blinder decomposition that parcels out racial and ethnic differences in the likelihood of hospitalization due to COVID-19 into three components (Jann 2008). The first component captures how much of the gap is from group differences in the predictors (the "endowments"). The second component captures the part due to differences in the coefficients (the "return to endowments"). The third component is the interaction between endowments and returns to endowments.

Furthermore, we perform a "detailed" decomposition, as we track two sets of predictors: comorbidities, which are indicators for the medical conditions we measure, and socio-demographic factors.[7] The decomposition is formulated from the viewpoint of Black or Hispanic enrollees. Our approach is designed to produce two relevant counterfactuals. First, what is the expected change in hospitalization rates for Black and Hispanic Americans if the relevant minority group had the majority group's predictor levels? Second, what would be the expected change if the minority group had the majority group's coefficients?[8]

Table 1 presents the decomposition results. The left panel displays results for non-Hispanic Black versus non-Hispanic White enrollees and the right panel displays results for Hispanic versus non-Hispanic White enrollees. The hospitalization rate for COVID-19 in our constructed sample is 7 percentage points higher for Black than White enrollees and 4.6 percentage points higher for Hispanic that White enrollees. For both groups, coefficients contribute much more to the overall difference than predictors. Perhaps surprisingly, the presence of comorbidities explains a much smaller share of the overall difference than the return to sociodemographic factors for both Black and Hispanic enrollees. Indeed, the return to sociodemographic factors is the single largest contributor to the overall gaps in hospitalization, accounting for 28.6 percent of the raw gap for Black compared to White enrollees and 56.7 percent of the raw gap for Hispanic compared to White enrollees.[9] The results indicate that the same predictors exert a more damaging impact on Black and Hispanic sampled enrollees. As an example, Black or Hispanic male enrollees might be more likely than White male enrollees in similar situations to be engaged in work-related activities that place them at higher risk of contracting

[7]For categorical variables, such as census block educational attainment and census division, we follow Yun (2005) to normalize the effects of categorical variables to avoid the issue of varying coefficients due to the choice of omitted group.

[8]We estimate with a linear probability model following Montenovo et al. (2020), but note that a logistic model provides similar results in terms of the importance of coefficients. Also, it is well-known that the reverse decomposition from the White perspective may provide different results, but our choice of perspective is shaped by the counterfactuals we wish to examine. See the Data Appendix for additional data and methodological details.

[9]The difference in intercept contributes negatively to the differences in returns to sociodemographic variables.

Table 1
Decomposition of Race-Based Differentials in Likelihood of Hospitalization Due to COVID-19

	Black versus White		*Hispanic versus White*	
Overall gap in sample	0.070		0.046	
	Comorbidities	*Sociodemographics*	*Comorbidities*	*Sociodemographics*
Endowments	0.011	0.007	0.003	0.001
Percent of total difference	16.2%	10.1%	6.5%	1.9%
Returns to endowments	0.016	0.020	0.012	0.026
Percent of total difference	22.8%	28.6%	24.9%	56.7%
Interaction	0.009	0.007	0.002	0.003
Percent of total difference	12.4%	9.9%	4.0%	6.1%
Number of observations	371,483		382,425	

Source: Authors' calculations from Optum (2021).
Note: Table reports results from a threefold Oaxaca-Blinder decomposition from the perspective of the minority group. Sociodemographics includes age and age squared, gender, education dummies, and census division fixed effects. Effects of education and census division are normalized. Comorbidities represent a series of dummy variables for hypertension, diabetes, obesity, heart disease, cancer and chronic obstructive pulmonary disease. The difference in intercepts is included in the difference in returns to endowments of sociodemographics. For details, see online Appendix available with this article at the *JEP* website.

the disease and/or have less access to care and therefore present at a later stage, thus requiring hospitalization.

In sum, the decomposition results suggest that the stark differences in COVID-19 health outcomes for Black and Hispanic Americans compared to non-Hispanic Whites cannot be attributed only to a greater prevalence of preexisting conditions, lower neighborhood levels of educational attainment, or (broad) geographical disadvantage. Rather, otherwise similar Black and Hispanic individuals, all of whom are insured in our sample, are hospitalized due to COVID-19 at a higher rate than non-Hispanic Whites. These results are specific to our sample and decomposition decisions we have taken, but they are consistent with the broader narrative that Black and Hispanic individuals face institutional disadvantages including inconsistent providers, lower-quality care, and systemic racism, that worsen their returns to similar endowments and contribute to COVID-19 health inequality.

Conclusion

This paper has drawn on history, medicine, and economics to place the initial health effects of the current pandemic into broader context. That COVID-19 disproportionately killed the frail and disadvantaged could be expected based on viral dynamics, past epidemics, and marked differences in the ability of individuals

to protect their health during the crisis. Yet the heightened salience of these patterns, along with the stark mortality burden of the pandemic, may serve as a catalyst for change—in particular, for changing how Americans view the importance of public health and the social safety net (Rees-Jones et al. 2020).

The initial health effects we outlined may have consequences for years to come. Most directly, a growing body of evidence suggests that a substantial proportion of individuals infected with COVID-19 suffer a range of long-term health consequences, including cognitive dysfunction, fatigue, and injury to the heart and lungs (COMEBAC Study Group 2021; del Rio, Collins, and Malani 2020). The pandemic's long-term effects, however, will likely reach past those who contracted the disease and extend beyond health alone. Disruptions in screenings and routine health care may beget future premature morbidity and mortality from other communicable and non-communicable diseases (Chen and McGeorge 2020). Scarring in utero exhibited in the influenza pandemic of 1918 and other epidemics of infectious disease may emerge with consequences for disability, educational attainment, and earnings (Almond 2006). The disruptive effects of the COVID-19 crisis on education could widen inequality in income and health for future generations. These factors may exacerbate immediate economic disparities already experienced through labor markets as a consequence of the pandemic (Montenovo et al. 2020).

With the introduction of recently approved vaccines against COVID-19 has come hope that the disruption the disease has wrought on health and society will soon cease. The same health gradients seen during the country's descent into the pandemic, however, are likely to be observed as we emerge from it. Reports suggest that vaccination distribution by race and ethnicity has not been aligned with who has been affected most by the virus, placing vulnerable individuals at risk of adverse outcomes during a time in which SARS-CoV-2 continues to mutate (Ndugga et al. 2021). The medium and long-run health effects of COVID-19—as well as the consequences of future novel infectious disease outbreaks which will assuredly emerge—will be shaped by how effectively and equitably policymakers respond to these formidable, yet not wholly unprecedented, challenges.

■ *We thank Cong Gian, Joyce Kim, and Nikhil Shankar for excellent research assistance. Mary Bassett, David Cutler, Ryan Edwards, Ronald Lee, Maria Polyakova, and Jonathan Skinner provided helpful comments. We thank Erik Hurst, Heidi Williams, and Timothy Taylor for very useful comments and suggestions on a previous draft of the paper.*

References

Abelson, Reed. 2021. "Buoyed by federal covid aid, big hospital chains buy up competitors." *The New York Times*, May 21. https://www.nytimes.com/2021/05/21/health/covid-bailout-hospital-merger.html.

Ahmad, Farida B., Jodi A. Cisewski, Arialdi Miniño, and Robert N. Anderson. 2021. "Provisional Mortality Data — United States, 2020." *Morbidity and Mortality Weekly Report* 70 (14): 519–22.

Almond, Douglas. 2006. "Is the 1918 Influenza Pandemic Over? Long-Term Effects of *In Utero* Influenza Exposure in the Post-1940 US Population." *Journal of Political Economy* 114 (4): 672–712.

Alsan, Marcella, and Sarah Eichmeyer. 2021. "Experimental Evidence on the Effectiveness of Non-Experts for Improving Vaccine Demand." NBER Working Paper 28593.

Alsan, Marcella, Owen Garrick, and Grant Graziani. 2019. "Does Diversity Matter for Health? Experimental Evidence from Oakland." *American Economic Review* 109 (12): 4071–4111.

Alsan, Marcella, Stefanie Stantcheva, David Yang, and David Cutler. 2020a. "Disparities in Coron- avirus 2019 Reported Incidence, Knowledge, and Behavior among Us Adults." *JAMA network open* 3 (6).

Alsan, Marcella M., Fatima Cody Stanford, Abhijit Banerjee, Emily Breza, Arun G. Chandrasekhar, Sarah Eichmeyer, Paul Goldsmith-Pinkham et al. 2020b. "Comparison of Knowledge and Information Seeking Behavior after General Covid-19 Public Health Messages and Messages Tailored for Black and Latinx Communities: A Randomized Controlled Trial." *Annals of Internal Medicine*.

Anderson, Kelly E., Emma E. McGinty, Rachel Presskreischer, and Colleen L. Barry. 2021. "Re- ports of Forgone Medical Care among US Adults during the Initial Phase of the Covid-19 Pandemic." *JAMA Network Open* 4 (1).

Andrasfay, Theresa, and Noreen Goldman. 2021. "Reductions in 2020 US Life Expectancy Due to Covid-19 and the Disproportionate Impact on the Black and Latino Populations." *Proceedings of the National Academy of Sciences of the United States of America* 118 (5).

Arrazola, Jessica, Matthew M. Masiello, Adrian E. Dominguez, Sujata Joshi, Amy Poel, Crisan- dra M. Wilkie, Jonathan M. Bressler et al. 2020. "Covid-19 Mortality among AmericanIndian and Alaska Native Persons — 14 States, January–June 2020." *Morbidity and Mortality Weekly Report* 69 (49): 1853–56.

Basu, Sanjay, Seth A. Berkowitz, Robert L. Phillips, Asaf Bitton, Bruce E. Landon, and Russell S. Phillips. 2019. "Association of Primary Care Physician Supply with Population Mortality in the United States, 2005–2015." *JAMA Internal Medicine* 179 (4): 506–14.

Birkmeyer, John D., Amber Barnato, Nancy Birkmeyer, Robert Bessler, and Jonathan Skinner. 2020. "The Impact of the Covid-19 Pandemic on Hospital Admissions in the United States." *Health Affairs* 39 (11).

Blau, Francine D., Josefine Koebe, and Pamela A. Meyerhofer. 2020. "Who Are the Essential and Front-Line Workers?" NBER Working Paper 27791.

Bushée, Frederick A. 1899. "The Growth of the Population of Boston." *Publications of the American Statistical Association* 6 (46): 239–74.

Centers for Disease Control and Prevention. 2020. *National Diabetes Statistics Report, 2020*. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services.

Centers for Disease Control and Prevention. 2021a. "2020–2021 Influenza Season for Week 21, Ending June 19, 2021." https://www.cdc.gov/flu/weekly/index.htm.

Centers for Disease Control and Prevention. 2021b. "Science Brief: Evidence Used to Update the List of Underlying Medical Conditions that Increase a Person's Risk of Severe Illness from Covid-19." https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/underlying-evidence-table.html.

Centers for Disease Control and Prevention. 2021c. "Trends in Number of Covid-19 Cases and Deaths in the US Reported to CDC, by State/Territory." https://covid.cdc.gov/covid-data-tracker/#trends_dailytrendscases.

Centers for Disease Control and Prevention, COVID-19 Response. 2020. "Covid-19 Case Surveillance Public Data Access, Summary, and Limitations (version date: December 31, 2020)." https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf. (accessed January 4, 2021).

Centers for Medicare and Medicaid Services. 2020. "CMS Releases Recommendations on Adult Elective Surgeries, Non-Essential Medical, Surgical, and Dental Procedures during Covid-19 Response."

https://www.cms.gov/newsroom/press-releases/cms-releases-recommendations-adult-elective-surgeries-non-essential-medical-surgical-and-dental.

Chandra, Amitabh, Pragya Kakani, and Adam Sacarny. 2020. "Hospital Allocation and Racial Disparities in Health Care." NBER Working Paper 28018.

Chen, Julius, and Rebecca McGeorge. 2020. "Spillover Effects of the Covid-19 Pandemic Could Drive Long-Term Health Consequences for Non-Covid-19 Patients." *Health Affairs Blog*, October 23. https://www.healthaffairs.org/do/10.1377/hblog20201020.566558/full/.

Cicala, Steve, Stephen P. Holland, Erin T. Mansur, Nicholas Z. Muller, and Andrew J. Yates. 2020. "Expected Health Effects of Reduced Air Pollution from Covid-19 Social Distancing." NBER Working Paper 27135.

Clark, Henry G., Charles E. Buckingham, John C. Dalton, and Henry W. Williams. 1850. *Report on the Cholera in Boston, in 1849.* Boston: J.H. Eastburn, City Printer.

Cohen, Robin A., Amy E. Cha, Michael E. Martinez, and Emily P. Terlizzi. 2020. "Health Insurance Coverage: Early Release of Estimates from the National Health Interview Survey, 2019." https://www.cdc.gov/nchs/data/nhis/earlyrelease/insur202102-508.pdf.

Cohn, D'Evra, and Jeffrey S. Passel. 2018. "A Record 64 Million Americans Live in Multigenerational Households." *Pew Research Center,* April 5. https://www.pewresearch.org/fact-tank/2018/04/05/a-record-64-million-americans-live-in-multigenerational-households/.

COMEBAC Study Group. 2021. "Four-Month Clinical Status of a Cohort of Patients after Hospitalization for Covid-19." *Journal of the American Medical Association* 325 (15): 1525–34.

Curtis, Josiah. 1856. *Report of the Joint Special Committee on the Census of Boston, May, 1855, Including the Report of the Censors, with Analytical and Sanitary Observations.* Boston: Moore & Crosby, City Printers—1 Water Street.

Czeisler, Mark E., Rashon I. Lane, Emiko Petrosky, Joshua F. Wiley, Aleta Christensen, Rashid Njai, Matthew D. Weaver et al. 2020a. "Mental Health, Substance Use, and Suicidal Ideation during the Covid-19 Pandemic — United States, June 24–30, 2020." *Morbidity and Mortality Weekly Report* 69 (32): 1049–57.

Czeisler, Mark E., Kristy Marynak, Kristie E.N. Clarke, Zainab Salah, Iju Shakya, JoAnn M. Thierry, Nida Ali et al. 2020b. "Delay or Avoidance of Medical Care because of Covid-19–Related Concerns — United States, June 2020." *Morbidity and Mortality Weekly Report* 69 (36): 1250–57.

Del Rio, Carlos, Lauren F. Collins, and Preeti Malani. 2020. "Long-Term Health Consequences of Covid-19." *Journal of the American Medical Association* 324 (17): 1723–24.

Doyle, Joseph J., John A. Graves, and Jonathan Gruber. 2019. "Evaluating Measures of Hospital Quality: Evidence from Ambulance Referral Patterns." *Review of Economics and Statistics* 101 (5): 841–52.

Funk, Cary, and Alec Tyson. 2020. "Intent to Get a COVID-19 Vaccine Rises to 60% as Confidence in Research and Development Process Increases." *Pew Research Center,* December 3. https://www.pewresearch.org/science/2020/12/ 03/intent-to-get-a-covid-19-vaccine-rises-to-60-as-confidence-in-research-and-development- process-increases/.

Gardner, John W., and Jill S. Sanborn. 1990. "Years of Potential Life Lost (YPLL) - What Does It Measure?" *Epidemiology* 1 (4): 322–29.

Gaynor, Martin. 2018. "Examining the Impact of Health Care Consolidation: Statement before the Committee on Energy and Commerce Oversight and Investigations Subcommittee, US House of Representatives." Unpublished.

Gichangi, Anthony, and Werner Vach. 2005. "The Analysis of Competing Risks Data: A Guided Tour." *Statistics in Medicine* 132: 1–41.

Goldstein, Joshua R., and Ronald D. Lee. 2020. "Demographic Perspectives on the Mortality of Covid-19 and Other Epidemics." *Proceedings of the National Academy of Sciences of the United States of America* 117 (36): 22035–41.

Gupta, Aakriti, Mahesh V. Madhavan, Kartik Sehgal, Nandini Nair, Shiwani Mahajan, Tejasav S. Sehrawat, Behnood Bikdeli et al. 2020. "Extrapulmonary Manifestations of Covid-19." *Nature Medicine* 26: 1017–32.

Halloran, M. Elizabeth. 2009. "Communicable Diseases and Data Analysis." *Biometrics* 2: 148–161.

Heist, Tyler, Karyn Schwartz, and Sam Butler. 2021. "Trends in Overall and Non-Covid-19 Hospital Admissions." *Kaiser Family Foundation*.

Holshue, Michelle L., Chas DeBolt, Scott Lindquist, Kathy H. Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters et al. 2020. "First Case of 2019 Novel Coronavirus in the United States." *New England Journal of Medicine* 382: 929–36.

Jann, Ben. 2008. "The Blinder–Oaxaca Decomposition for Linear Regression Models." *The Stata Journal* 8: 453–79.

Jha, Ashish K., E. John Orav, and Arnold M. Epstein. 2011. "Low-Quality, High-Cost Hospitals, Mainly in South, Care for Sharply Higher Shares of Elderly Black, Hispanic, and Medicaid Patients." *Health Affairs* 30 (10): 1904–11.

Jose, Ricardo J., and Ari Manuel. 2020. "Covid-19 Cytokine Storm: The Interplay between Inflammation and Coagulation." *Lancet Respiratory Medicine* 8: E46–47.

Kakani, Pragya, Amitabh Chandra, Sendhil Mullainathan, and Ziad Obermeyer. 2020. "Allocation of Covid-19 Relief Funding to Disproportionately Black Counties." *Journal of the American Medical Association* 324 (10): 1000–1003.

Krishnan, Lakshmi, S. Michelle Ogunwole, and Lisa A. Cooper. 2020. "Historical Insights on Coronavirus Disease 2019 (Covid-19), the 1918 Influenza Pandemic, and Racial Disparities: Illuminating a Path Forward." *Annals of Internal Medicine* 173 (6): 474–82.

Long, Brit, William J. Brady, Alex Koyfman, and Michael Gottlieb. 2020. "Cardiovascular Complications in Covid-19." *American Journal of Emergency Medicine* 38 (7): 1504–07.

Malas, Mahmoud B., Isaac N. Naazie, Nadin Elsayed, Asma Mathlouthi, Rebecca Marmor, and Bryan Clary. 2020. "Thromboembolism Risk of Covid-19 Is High and Associated with a Higher Risk of Mortality: A Systematic Review and Meta-Analysis." *Lancet EClinicalMedicine* 29–30.

Mein, Stephen A. 2020. "Covid-19 and Health Disparities: The Reality of 'The Great Equalizer'." *Journal of General Internal Medicine* 35 (8): 2439–40.

Montenovo, Laura, Xuan Jiang, Felipe Lozano Rojas, Ian M. Schmutte, Kosali I. Simon, Bruce A. Weinberg, and Coady Wing. 2020. "Determinants of Disparities in Covid-19 Job Losses." NBER Working Paper 27132.

National Center for Health Statistics. 2020a. "Bridged-Race Population Estimates 1990–2002." https://wonder.cdc.gov/Bridged-Race-v2002.HTML. (accessed May 11, 2021).

National Center for Health Statistics. 2020b. "Bridged-Race Population Estimates 1990–2019." https://wonder.cdc.gov/Bridged-Race-v2019.HTML. (accessed May 11, 2021).

National Center for Health Statistics. 2020c. "Compressed Mortality File: 1979–1998 with icd-9 codes." https://wonder.cdc.gov/cmf-icd9.html. (accessed January 11, 2021).

National Center for Health Statistics. 2020d. *Excess Deaths Associated with Covid-19*. Washington, DC: Center for Disease Control and Prevention.

National Center for Health Statistics. 2020e. "Underlying Cause of Death, 1999–2019." https://wonder.cdc.gov/ucd-icd10.html. (accessed May 11, 2021).

National Center for Health Statistics. 2021a. "AH Monthly Provisional Counts of Deaths for Select Causes of Death by Age, and Race and Hispanic Origin." https://data.cdc.gov/NCHS/AH- Monthly-Provisional-Counts-of-Deaths-for-Select/r5pw-bk5t.

National Center for Health Statistics. 2021b. "Provisional Drug Overdose Death Counts." https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm.

National Center for Health Statistics. 2021c. "Technical Notes: Provisional Death Counts for Coronavirus Disease (Covid-19)." https://www.cdc.gov/nchs/nvss/vsrr/covid19/tech_notes.htm.

National Research Council. 2009. *Vital Statistics: Summary of a Workshop*. Washington, DC: National Academies Press.

Ndugga, Nambi, Olivia Pham, Latoya Hill, Samantha Artiga, Raisa Alam, and Noah Parker. 2021. "Latest Data on Covid-19 Vaccinations Race/Ethnicity." *Kaiser Family Foundation*, June 16. https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-race-ethnicity/.

Optum. 2021. "Clinformatics® Data Mart." (accessed May 11, 2021).

Patel, Sadiq Y., Ateev Mehrotra, Haiden A. Huskamp, Lori Uscher-Pines, Ishani Ganguli, and Michael L. Barnett. 2021. "Trends in Outpatient Care Delivery and Telemedicine during the Covid-19 Pandemic in the US." *JAMA Internal Medicine* 181 (3): 388–91.

Petrilli, Christopher M., Simon A. Jones, Jie Yang, Harish Rajagopalan, Luke O'Donnell, Yelena Chernyak, Katie A. Tobin, Robert J. Cerfolio, Fritz Francois, and Leora I. Horwitz. 2020. "Factors Associated with Hospital Admission and Critical Illness among 5279 People with Coronavirus Disease 2019 in New York City: Prospective Cohort Study." *BMJ* 369.

Physicians Foundation. 2020. "2020 Survey of America's Physicians." https://physiciansfoundation.org/research-insights/2020physiciansurvey/.

Polyakova, Maria, Geoffrey Kocks, Victoria Udalova, and Amy Finkelstein. 2020. "Initial Economic Damage from the COVID-19 Pandemic in the United States Is More Widespread across Ages and

Geographies than Initial Mortality Impacts." *Proceedings of the National Academy of Sciences of the United States of America* 117 (45): 27934–39.

Polyakova, Maria, Victoria Udalova, Geoffrey Kocks, Katie Genadek, Keith Finlay, and Amy Finkelstein. 2021. "Racial Disparities in Excess All-Cause Mortality during the Early COVID-19 Pandemic Varied Substantially across States." *Health Affairs* 40 (2).

Rees-Jones, Alex, John D'Attoma, Amedeo Piolatto, and Luca Salvadori. 2020. "COVID-19 Changed Tastes for Safety-Net Programs." NBER Working Paper 27865.

Reeves, Richard V., and Faith Smith. 2020. "Black and Hispanic Americans at Higher Risk of Hypertension, Diabetes, Obesity: Time to Fix Our Broken Food System." *Brookings Up Front*, August 7. https://www.brookings.edu/blog/up-front/2020/08/07/black-and-hispanic-americans- at-higher-risk-of-hypertension-diabetes-obesity-time-to-fix-our-broken-food-system/.

Rubin-Miller, Lily, Christopher Alban, Samantha Artiga, and Sean Sullivan. 2020. "COVID-19 Racial Disparities in Testing, Infection, Hospitalization, and Death: Analysis of Epic Patient Data." *Kaiser Family Foundation*, September 16. https://www.kff.org/coronavirus-covid-19/issue-brief/covid-19-racial-disparities-testing-infection-hospitalization-death-analysis-epic-patient-data/.

Ruhm, Christopher J. 2000. "Are Recessions Good for Your Health?" *Quarterly Journal of Economics* 115 (2): 617–50.

Shattuck, Lemuel. 1846. *Report to the Committee of the City Council Appointed to Obtain the Census of Boston for the Year 1845: Embracing Collateral Facts and Statistical Researches, Illustrating the History and Condition of the Population, and Their Means of Progress and Prosperity.* Boston: J.H. Eastburn, City Printer.

Simonds, Artemas. 1850. *Report by the City Registrar of the Births, Marriages, and Deaths, in the City of Boston, for the Year 1849.* Boston: J.H. Eastburn City Printer.

Simonov, Andrey, Szymon K. Sacher, Jean-Pierre H. Dubé, and Shirsho Biswas. 2020. "The Persuasive Effect of Fox News: Non-Compliance with Social Distancing during the COVID-19 Pandemic." NBER Working Paper 27237.

Snowden, Lonnie R., and Genevieve Graaf. 2021. "COVID-19, Social Determinants Past, Present, and Future, and African Americans' Health." *Journal of Racial and Ethnic Health Disparities* 8: 12–20.

Stevens, Ann H., Douglas L. Miller, Marianne E. Page, and Mateusz Filipski. 2015. "The Best of Times, the Worst of Times: Understanding Pro-cyclical Mortality." *American Economic Journal: Economic Policy* 7 (4): 279–31.

US Census Bureau. 1919. "Mortality Statistics 1917, eighteenth annual report." https://www.cdc.gov/nchs/data/vsushistorical/mortstatsh_1917.pdf. (accessed January 28, 2021).

US Census Bureau. 1920. "Mortality Statistics 1918, nineteenth annual report." https://www.cdc.gov/nchs/data/vsushistorical/mortstatsh_1918.pdf. (accessed January 28, 2021).

US Census Bureau of the Census. 1922. "1920 Census: Volume 2. population, general report, and analytical tables, chapter 3." https://www2.census.gov/library/publications/decennial/1920/volume-2/41084484v2ch03.pdf. (accessed January 28, 2021).

US Census Bureau. 2020a. "2010–19 American Community Survey 1-Year Estimates, Tables b01001b, b01001c, b01001d, b01001h, and b01001i." https://data.census.gov/cedsci/. (accessed May 11, 2021).

US Census Bureau. 2020b. "Household Pulse Survey Public Use File (puf)." https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html. (accessed January 13, 2021).

Virchow, Rudolf. 1848. *Die Medizinische Reform. Eine Wochenschrift*, 1, 182.

Whaley, Christopher M., Megan F. Pera, Jonathan Cantor, Jennie Chang, Julia Velasco, Heather K. Hagg, Neeraj Sood, and Dena M. Bravata. 2020. "Changes in Health Services Use among Commercially Insured US Populations during the COVID-19 Pandemic." *JAMA Network Open* 3 (11).

Wright, Ephraim M. 1855. *Thirteenth Report to the Legislature of Massachusetts, Relating to the Registry and Returns of Births, Marriages, and Deaths, in the Commonwealth, for the Year Ending December 31, 1854.* Boston: William White, Printer to the State.

Wu, Sean L., Andrew N. Mertens, Yoshika S. Crider, Anna Nyugen, Nolan N. Pokpongkiat, Stephanie Djajadi, Anmol Seth et al. 2020. "Substantial Underestimation of SARS-CoV-2 Infection in the United States." *Nature Communications* 11.

Yellow Horse, Aggie J., and Kimberly R. Huyser. 2021. "Indigenous Data Sovereignty and COVID-19 Data Issues for American Indian and Alaska Native Tribes and Populations." *Journal of Population Research.* Yun, Myeong-Su. 2005. "A Simple Solution to the Identification Problem in Detailed Wage Decompositions." *Economic Inquiry* 43 (4): 766–72.

Ziedan, Engy, Kosali Simon, and Coady Wing. 2020. "Effects of State COVID-19 Closure Policy on Non-COVID-19 Health Care Utilization." NBER Working Paper 27621.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 47–66

Tracking the Pandemic in Real Time: Administrative Micro Data in Business Cycles Enters the Spotlight

Joseph Vavra

Looking back, we now know that the US unemployment rate at the start of the COVID-19 pandemic rose from 3.2 percent in February 2020 to 4.1 percent in March and 13.1 percent in April. However, the April unemployment rate was not reported by the Bureau of Labor Statistics until early May. Preliminary data on April retail sales was not released by the Census Bureau until mid-May, and the first release of gross domestic product data by the Department of Commerce covering April did not occur until the end of July. Thus, a number of economists turned to private-sector micro data to try to understand the recession while it was still unfolding: for example, data on employment patterns from the payroll processing firm ADP and the scheduling firm Homebase, data on bank accounts and credit card payments from sources like the JPMorgan Chase Institute and firms that provide financial planning services like mint.com and SaverLife, and even data on locations of cell phone users from firms like PlaceIQ and SafeGraph. The use of administrative micro data from these and other sources allowed pandemic-related research to be produced in nearly real-time and the scope for analysis of individual behavior, which would be impossible using traditional aggregate data.

In this essay, I loosely define administrative data as that arising as a by-product of some non-research activity, which contrasts with traditional data sources that are primarily collected for research purposes, like the Panel Survey of Income Dynamics, the American Community Survey, or the Consumer Expenditure Survey. The applications I discuss in this paper use administrative data collected by

■ *Joseph Vavra is Professor of Economics, Booth School of Business, University of Chicago, Chicago, Illinois. His email address is joseph.vavra@chicagobooth.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.47.

mortgage servicers, cell phone apps, credit bureaus, financial services firms, and payroll processing firms. These companies collect vast amounts of data in the course of their regular business, but this data is not collected with academic research as an end-goal. While this definition of administrative data could also include data produced by government administrations as a product of non-research activity, such as micro data on taxes and households from the Internal Revenue Service or data based on information from state-level unemployment insurance agencies like the Longitudinal Household-Employment Dynamics data, I focus my discussion on privately collected data.

Over the last decade, there has been an explosion in the availability and use of administrative micro data for economic research—this trend has cut across most empirical subfields of economics. But in this paper, I discuss ways in which this data has shaped macroeconomic research on recessions and stimulus policy. The Great Recession was the first business cycle to occur in this new age of administrative data availability, and although the research was using this data retrospectively, I begin the paper with a brief discussion of some macro lessons from this research. However, I focus mostly on the pandemic, because administrative data has played a crucial, early role in shaping our understanding of this period. For example, the massive and incredibly rapid increases in unemployment at the start of the pandemic were particularly concentrated in low income, service-sector jobs, while high income workers were largely insulated from job loss and saw only modest cuts in nominal wages. Expanded unemployment benefits had a substantial and immediate effect on spending, but did little to discourage job search.

In some cases, the patterns uncovered with studies of administrative data can also be seen with traditional data sources, but they were apparent weeks or months sooner because administrative data is often available in nearly real-time. This opens up the possibility for faster policy reactions. In other cases, administrative data leads to insights which cannot be obtained with traditional data sources. Traditional survey-based data typically either have small sample sizes or a limited panel element, they often have non-trivial measurement error, and they are often released with substantial lags. Administrative data can provide novel insights by measuring variables with more precision than traditional data methods or by measuring variables that are not captured by any traditional data sources. In addition, vast sample sizes can enable very detailed data cuts and statistical precision and can sometimes allow for new sources of variation and identification strategies.

Administrative data also raises challenges and concerns. The raw data itself was collected for other purposes, and for researchers, it can often be messy and difficult to interpret. The nature of administrative data means it often has a narrow lens of focus, with great depth but more limited breadth. Representativeness and external validity of administrative data are often big concerns. The greater statistical precision does not necessarily mean that estimates are unbiased, because large sample sizes do not themselves solve identification challenges and resolve issues of causality.

I do not attempt to provide a comprehensive review of all research that uses administrative micro data to understand recessions; indeed, given the explosion of

work in this area and the many administrative datasets now in use, that task would be enormous indeed. Instead, I intentionally choose a small number of applications, and only a few papers within each application to illustrate some of the breadth of administrative micro data available as well as some common challenges. I focus on applications to the US macroeconomy, but there is clearly also a vast amount of administrative micro data, policy variation, and applications of interest in other countries.

Administrative Data in the Great Recession

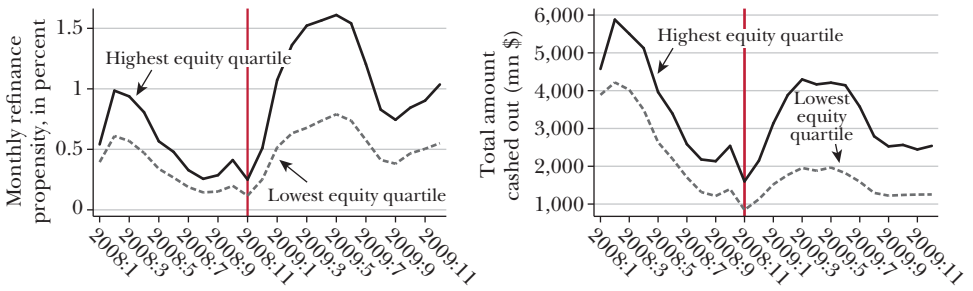
The Great Recession from 2007 to 2009 offered the first widespread application of administrative micro data to understanding recessions. The key role of house price declines and mortgage market disruptions in the Great Recession is widely established, most prominently with the work of Mian and Sufi (for an overview, see their 2018 essay in this journal). The fact that the recession itself originated in mortgage markets made data on mortgages (and credit more broadly) vital for understanding this period, and the simultaneous rise in "big data" information technology infrastructure meant that credit providers had such data. Many lenders were willing to make this data available to researchers, and macroeconomists took advantage. Here, I focus on some applications from my own work to illustrate some of the insights and the limits of such data.

In Beraja et al. (2019), we use administrative micro data to show that in addition to this direct effect of house prices on the economy, the house price boom–bust also substantially constrained the monetary policy response to the Great Recession. Collateralized borrowing in the housing market is an important part of the monetary transmission mechanism, because interest rate cuts encourage households to refinance their mortgage and extract home equity to fund current consumption. However, using the Credit Risk Insight Servicing McDash (CRISM) dataset produced by merging mortgage servicing records from Black Knight Financial Services (BKFS) with credit bureau data from Equifax, we show that this transmission mechanism was substantially dampened during the Great Recession.

This monthly panel data includes detailed loan-level characteristics including loan balances, interest rates, and origination characteristics for tens of millions of loans which are serviced through BKFS. These loans are, in turn, linked to Equifax borrower credit records, allowing us to measure consolidated borrowing positions, creditworthiness, and, most importantly, letting us link successive loans for a specific borrower across time. This monthly data ultimately allows us to measure refinancing and home equity extraction after the large declines in mortgage rates induced by the first round of the Federal Reserve's quantitative easing program in November 2008.

During the Great Recession, house prices fell substantially on average, but declines varied greatly across space. Using the CRISM micro data, we show that interest rate declines during the Great Recession had the smallest effects on

Figure 1
Mortgage Refinancing and Cash-Outs during the Great Recession



Source: This figure is reproduced from Beraja et al. (2019, Figures 3a and 4a).
Note: Left panel shows refinancing propensities in the lowest equity (largest house price decline) compared to highest equity metropolitan statistical areas, and the right panel shows cash-out volumes.

refinancing in the locations with the largest house price declines and increases in unemployment. For example, both refinancing in general and cash-out activity, in particular, rose much less after quantitative easing (marked by the red lines in Figure 1) in these locations with little housing equity. An obvious explanation is that many households were underwater in these locations, making it difficult or impossible to refinance, and for households who could refinance, they had more limited equity to extract.

Translating this cross-region evidence through the lens of a macro model, we conclude that the house-price bust substantially constrained the stimulative power of monetary policy during the Great Recession. In closely related work, Berger et al. (forthcoming) uses this same CRISM data over a much longer period of time to show that the strength of this refinancing channel of monetary policy is influenced not just by house price movements but also by the past behavior of monetary policy itself. Keeping interest rates low for an extended period of time, as has been done since the Great Recession, means that many households lock in low interest rates and become less sensitive to future rate stimulus—even after rates return to more normal levels.

This research using administrative data provides a new basis for thinking about the aggregate strength of this refinancing channel. First, while refinancing has been studied empirically for almost 40 years, older research tended to focus on data with relatively small samples, which may or may not be representative when trying to draw inference for the economy as a whole. In contrast, the CRISM data covers around 60 percent of the entire mortgage market, with a broad cross-section of loan types and characteristics, so it is much more likely to be representative. Second, prior research typically studies loan-level datasets without links across time, which means earlier studies cannot separately distinguish loan prepayment arising from rate refinancing, cash-out refinancing, moving, and default.

However, this data is not a panacea, and it faces some limitations and pitfalls in answering some important questions. The data provides a very detailed lens into mortgage characteristics and household liabilities, but it has essentially no useful information on other important aspects of the household balance sheet like spending, assets, or income. Because this data has greater depth than breadth in its lens of coverage, it is difficult to answer important questions like how spending responds to mortgage refinancing or how refinancing responds to income shocks. While some papers (like Bartlett et al. 2019) have linked this data to other administrative datasets to expand the set of demographic information available, to my knowledge, no links to many key covariates like income exist.

On this particular point, it is useful to highlight a potential pitfall that arises frequently in the use of administrative micro data: imputed data. Casual users of Equifax CRISM data may misleadingly think that Equifax *does* collect information on income because their data reports such information. However, these variables are entirely imputed rather than coming from any actual data on income, so they are of little practical use. This particular data is well-documented and the imputation points are clear from inspection of data codebooks. However, the point illustrates a broader practical concern with the recent rise of various "big data" insight providers that mark micro data products. These products often draw from many different sources with little transparency and market the breadth of data that they provide—even though some of this data is imputed or predicted using machine learning. It is important, but not always possible, to distinguish actual from imputed data in these sources when commercial motives mean there is little transparency about underlying sources or methods. While this concern is clearly data-set specific, I recommend that researchers invest time to really understand the collection and construction of the data that they use in order to limit these potential problems.

While a lot of research is focused on mortgage markets, many papers also used administrative micro data to explore more typical macroeconomic questions during the Great Recession. For example, Stroebel and Vavra (2019) and Grigsby, Hurst, and Yildirmaz (2021) study the cyclicality of prices and nominal wages, respectively. Stroebel and Vavra (2019) use weekly store-UPC pricing data from the marketing firm IRI to construct local price indices and use these indices in a cross-zip-code-based identification strategy to argue for procyclical price markups. This type of analysis would be impossible with traditional price indices, which are only available at much higher levels of spatial aggregation. Grigsby, Hurst, and Yildirmaz (2021) use ADP data (discussed in more detail in the following section) to show that nominal wage cuts were much less common during the Great Recession than implied by traditional datasets. In traditional datasets, it is very difficult to measure the frequency and size of nominal wage adjustment because even tiny measurement errors can contaminate results. ADP data measures the actual wages paid using administrative data from those paying the wage and can thus eliminate wage "cuts" arising from measurement error.

All of the Great Recession research highlighted in this section exploits administrative micro data for research, which could not be performed with traditional

datasets. In general, this type of unique analysis is where administrative data has the highest value added. However, even in these research applications which would be impossible with traditional sources, it is important to highlight that administrative data does not supplant traditional data. This research still relies on traditional data sources for crucial benchmarking steps and validation of representativeness, which are pervasive concerns with administrative data.

Great Recession research demonstrated the fundamental value of administrative micro data for macroeconomics. Furthermore, the relationships established between academic researchers and data providers through this research, in turn, played a crucial role in speeding analysis of the pandemic when it arrived.

Administrative Data in the Pandemic

Economists have responded to the worldwide health crisis with an unusually rapid and focused outpouring of research on its economic effects. This analysis has been produced much more rapidly than in the Great Recession, often being released weeks or even just days after relevant events. This pace of research opens new opportunities for influencing policy as it unfolds rather than later analyzing the consequences of policy, but it also introduces a number of novel challenges. There are obvious trade-offs between producing deep and careful research that will stand the test of time and producing research quickly. Indeed, the fast pace of this research means that more findings will likely eventually need to be revised or clarified relative to research conducted at a more typical academic pace.

In my discussion here, I choose applications with two goals in mind: 1) I want to highlight several broadly different types of administrative data used to understand the pandemic recession; and 2) I want to highlight results and conclusions that have received some amount of support in multiple administrative data sources or with traditional data. Most of the research I mention here focuses on the period from March to September 2020. We know that underlying health and economic conditions have changed rapidly across time, so it is important to note that conclusions from research looking at this early stage of the pandemic may differ from research examining the current stage of the pandemic or the eventual recovery over the coming months or years.

Labor Market Data

Some of the first empirical research on the pandemic focused on measuring labor market disruptions using administrative micro data. Cajner et al. (2020) and Grigsby et al. (2021) use data from the payroll processing firm ADP, and Bartik et al. (2020) use data from the scheduling firm Homebase to document a number of labor market facts in the early stages of the pandemic.

ADP is a large human resources firm providing payroll processing for around 26 million US workers each month. This data is broadly representative of private sector employment using a variety of external benchmarks, although it modestly

overweights medium-large firms (Cajner et al. 2020; Grigsby, Hurst, and Yildirmaz 2021). Homebase is a scheduling firm providing services to tens of thousands of small businesses that employ hundreds of thousands of workers. This dataset is much less representative because it is skewed towards small firms in sectors like restaurants and retail that disproportionately employ hourly workers. However, these firms were among those most disrupted by the pandemic and they are otherwise somewhat underrepresented in the ADP data. Furthermore, Bartik et al. (2020) complement the raw data with an additional survey of Homebase users, allowing for some additional insights using this data.

Both papers show the striking distributional effects of the pandemic: lower-wage workers were much more likely to lose their jobs than higher-wage workers during this time period. Figure 2 illustrates these findings. In part, this pattern arises because low-wage workers tend to be concentrated in sectors like food service, which were particularly hard hit by the pandemic. Furthermore, this specific low-wage segment of the population is particularly vulnerable because these workers are also less likely to have substantial savings.

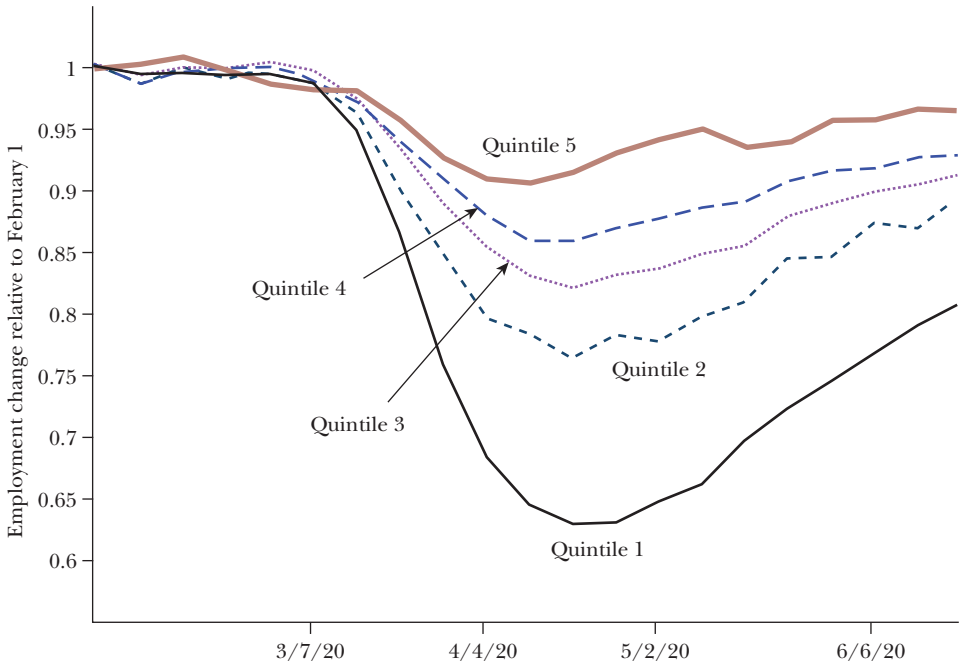
These broad distributional patterns are masked when focusing on the aggregate unemployment rate, and they were first established in these administrative datasets. Several papers have now documented this same pattern of greater unemployment for low-wage workers using traditional public datasets (for example, Ganong, Noel, and Vavra 2020; Cortes and Forsyth 2020). Thus, a main advantage of administrative data in this context was its speed, rather than a unique lens. Using administrative data to understand labor market trends 4-6 weeks earlier is of great use for policy-making decisions, but is arguably less crucial at the typical horizons of academic research.[1]

However, these administrative data studies also offered some more unique insights. Bartik et al. (2020) decompose the total reduction of worker hours and find it was primarily driven by extensive margin effects with firms shutting down entirely or reducing the size of their workforce, rather than intensive margin effects where hours were reduced but workers remained employed.

As discussed in the previous section, ADP data has a unique ability to measure nominal wage adjustment because it measures actual payments made to workers and thus does not suffer from measurement error, which contaminates traditional survey-based data. Following Grigsby, Hurst, and Yildirmaz (2021), Grigsby et al. (2021) find that 6 percent of workers (mostly at the top of the income distribution) received nominal wage cuts early in the pandemic, but that 30 percent of these wage cuts were reversed by November. The pace of this nominal wage adjustment

[1] Indeed, publicly available labor market data is itself already available quite rapidly. Thus, speed will generally be a greater comparative advantage for administrative data on spending, rather than for labor market data, since public data on spending is produced with moderately longer lags. See Chetty et al. (2020) for an effort to produce and publicly distribute daily statistics on consumer spending, business revenues, employment rates, and other key indicators disaggregated by ZIP code, industry, income group, and business size, based on anonymized data from a group of companies. For details, see https://tracktherecovery.org/.

Figure 2
Employment Change Relative to February 1, 2020, by Income Quintile



Note: Computed with ADP data in Cajner et al. (2020, their Figure 4a). The lines show pre-pandemic wage quintiles.

is substantially greater than during the Great Recession, but it still implies modest effects on overall earnings relative to the layoffs at the bottom end of the distribution. Overall, this data shows that the main labor market effect of the pandemic has been a large increase in unemployment at the bottom of the distribution, and that there is a more modest decline in wages but with continuing employment at the top of the distribution.

Moving forward, there will undoubtedly be much more research using this administrative data to understand labor markets. Two of the biggest advantages of these administrative data relative to traditional data sources are the ability to link individual workers together with firms so that workers can be tracked over time, and the fact that pay and hours can be measured exactly without the measurement error from self-reported data. On the other side, a potentially significant concern is that this data captures small employers (Homebase) or broader private-sector employment (ADP), but it has essentially no information on public-sector employment. If state and local budget cuts (early in the pandemic) or surpluses (later in the pandemic) lead to public sector employment changes, this data will largely miss these trends.

Financial Accounts Data

Data from individual bank accounts can be used to study various household-level outcomes. The first and most direct source of account-level data are banks themselves. The primary source of such data in the United States is the JPMorgan Chase Institute (JPMCI), a think tank within JPMorgan Chase & Co., which has developed a strictly controlled process to use anonymized account-level data on the universe of Chase customers directly for academic and policy research. The second common source of bank account data comes from financial service companies, which often require the user to first provide bank account log-in information to obtain some service; once the company obtains this data, they then make anonymized versions available for research purposes. For example, users of mint.com and SaverLife users enter all of their various account information, and these companies then provide financial planning services and budgeting information to their users. Users of Earnin can sign up to receive free payday loans, but they must first link to a bank account in order to do so.[2]

Bank account information provides a detailed and high-frequency lens into individual economic behavior. It typically contains transaction-level information on both account inflows (like direct deposits) and account outflows (like debit card transactions), allowing researchers to measure the connection among high frequency income, spending, and savings. Other datasets have detailed information on individual components (for example, the ADP data described in the previous section for income, or data produced by Visa or credit card processing companies for spending), but cannot link these components at a household level. Such links turn out to be crucial for some of the insights using this data to study the pandemic.

The JPMCI data has the further advantage of large sample sizes: as of 2015, it includes 27 million checking accounts (Farrell and Greig 2015). In addition, Cox et al. (2020) shows that the distribution of income in this data is generally similar to that of the population as a whole (although by construction it does not include any "unbanked" households). While the JPMCI data has essentially a complete, transaction-level accounting of everything that occurs within Chase accounts, a corresponding disadvantage of this data is that it has a limited lens for anything that occurs outside of Chase accounts, such as activities in second bank accounts or on non-Chase credit cards. For this reason, most research using JPMCI introduces various screens so that inactive or barely active accounts are not included in the analysis, but it is nevertheless possible that some important non-Chase activity is missed.

An advantage of bank account data from financial aggregators like SaverLife is that users have strong incentives to include all of their active accounts in order to obtain reliable planning information. These datasets do tend to have much smaller sample sizes, and selection is more of a concern because users choosing to

[2]Facteus also provides some similar information from card processing, combining individuals using debit cards, payroll cards, and load cards at an account level, although I have not worked with this data and am less clear on the underlying nature of the sample.

use financial planning apps may not be representative of the broader population. However, Baker (2018) provides substantial benchmarking and validation exercises to argue that these selection concerns are of more limited import in his applications and that data from these types of platforms are indeed informative about broader behavior. While few studies have used Earnin data, these selection issues are likely to be an even greater concern for that data because this sample is built on those seeking payday loans. However, that data may be useful for understanding the behavior of particular vulnerable populations of interest that might be underrepresented in other datasets.

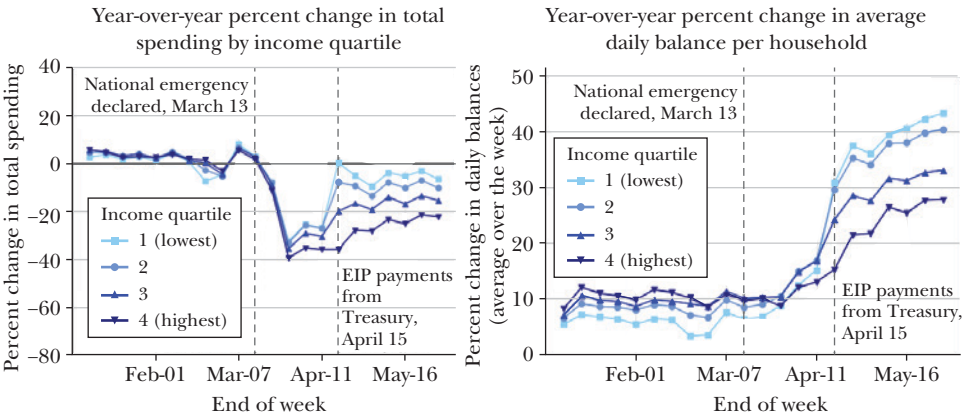
What was learned from administrative bank account data during the pandemic? Cox et al. (2020) complement the analysis of labor income losses discussed earlier by using JPMCI data to show how spending and savings have changed across the income distribution over the same period. This analysis requires linking individual income, spending, and savings, which is possible only as a result of the unique lens offered by administrative financial account data.[3] Using a sample of around five million active account-holders for which they can measure pre-pandemic income using direct deposit information, Cox et al. (2020) find dramatic declines in spending across the income distribution at the end of March 2020. These spending declines are strongest in certain categories like entertainment and hotel accommodations, which require in-person activity.[4] However, starting in mid-April, spending recovers much more rapidly for *low-income* households. At the same time, these low-income households also see the largest growth in checking account balances. Figure 3 illustrates these patterns. This finding seems surprising in light of the evidence from Cajner et al. (2020) and Bartik et al. (2020) that these households had the largest declines in labor income over this same period in time.

How can the households experiencing the most job loss during the pandemic fare best in terms of spending and savings growth? The timing suggests an important role for government support programs. The divergence in spending patterns occurred shortly after the passage of the Coronavirus Aid, Relief, and Economic Security (CARES) Act, which created large transfers that disproportionately benefited low-income households. In particular, the CARES Act provided one-time broad-based Economic Impact Payments of $1,200 for most adults and created a Federal Pandemic Unemployment Compensation program that added $600 per week on top of regular state unemployment insurance benefits from April through July. The Economic Impact Payments were the same absolute size for all but the

[3] Some of these spending patterns could potentially be explored in the publicly available consumer expenditure data once it becomes available covering this period. In addition, although Chetty et al. (2020) do not have individual income data, they find similar zip code–level spending patterns across zip codes with high and low income, suggesting that these spending patterns over the income distribution are not unique to JPMCI data. Finally, it is useful to note that the large run-up in savings observed in JPMCI is consistent with the spike in savings observed in aggregate data once it was released.

[4] These category spending patterns do not require account-level links and have been observed in a variety of credit card data sources including Affinity (Chetty et al. 2020) and Womply (Alexander and Karger 2020). These papers include a much more extensive discussion of these category patterns.

Figure 3
Changed in Total Spending and Average Daily Balances during the Pandemic



Source: Cox et al. (2020, Figure 5a and 11b).

highest-income individuals, so they resulted in larger income growth for low-income households. The $600 unemployment insurance supplements even more disproportionately benefited low-income households, in part because low-income households were more likely to be unemployed, and in part because a flat weekly $600 supplement represents a larger share of pre-job loss income for low income individuals. Figure 3 uses vertical lines to show the date when the national pandemic emergency was declared and when Economic Impact Payments were first distributed by the US Treasury.

These correlations with policy timing are merely suggestive, but several papers have used administrative data to make an argument for causal effects of these expanded transfers. Baker et al. (2020) use the SaverLife data described above to analyze high frequency spending responses to the one-time Economic Impact Payments. Using daily data on 38,000 active account users from December 2019 through May 2020 they identify just over 23,000 Economic Impact Payments by looking for direct and check deposits in certain categories with sizes corresponding to common amounts of these payments, like $1,200 and $2,400. They then estimate a marginal propensity to consume from 7 days before to 23 days after the receipt of the payments, using a distributed lag regression controlling for individual and time fixed effects to estimate responses. Overall, they find a cumulative marginal propensity to consume of around $0.37 in their raw sample, or $0.27 when they reweight their relatively low-income sample to instead match the distribution of income and several other observables in Current Population Survey data. Users with earnings under $1,000 per month had a marginal propensity to consume roughly twice as large as users earning $5,000 a month or more. There is an even steeper gradient with liquidity, as users with the highest account balances have marginal propensities to consume around 0.1 while those with balances under $100 have marginal

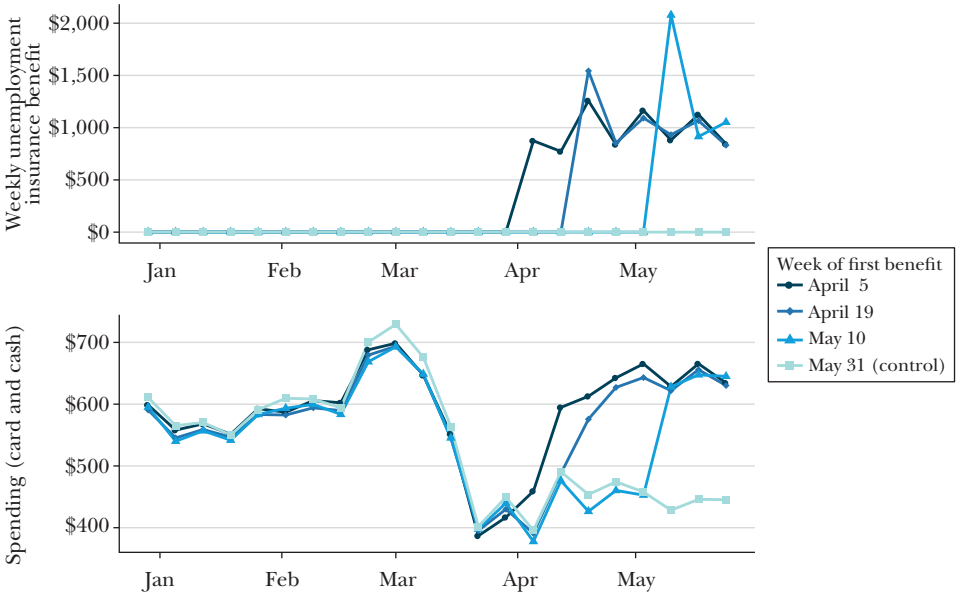
propensities to consume of 0.4 or more. Overall, these results are consistent with evidence from prior regressions in Johnson, Parker, and Souleles (2006) and Parker et al. (2013) that stimulus checks significantly boosted spending.

However, even though the SaverLife data is vastly superior in frequency, detail, and accuracy to the Consumer Expenditure Survey data used in studies of earlier recessions, the policy variation itself during the pandemic is harder to interpret from a causal identification perspective. Unlike in the 2001 and 2008 recessions, the timing of stimulus checks during the pandemic was not random and was concentrated over a shorter period. Paper checks during the pandemic were prioritized by income, and direct deposit timing depended on past tax filing status and differed for those on Social Security and other government support programs, all of which are correlated with income. The non-random timing of payments along with the differential trends in spending by income shown in Cox et al. (2020) introduce some concerns about interpreting their estimates of marginal propensity to consume, especially at longer horizons, as pure causal effects. I note this not as a critique of their general conclusion that stimulus checks increased spending. Given the very high frequency of sharp breaks in spending observed in Baker et al. (2020), there is little doubt that the stimulus checks did boost spending. But I do want to highlight the broader point that using administrative data, even when it might have enormous sample sizes and more precise measurement, does not in-and-of-itself solve identification challenges.

Overall, Economic Impact Payments were a large part of the initial US government response to the pandemic, totaling $270 billion by the end of May 2020, and the evidence in Baker et al. (2020) implies they had an important role in increasing spending, especially for lower-income households as shown in Figure 4. In addition to these broad-based payments, the more targeted $600 weekly supplements to unemployment insurance also played a particularly important role in increasing spending for low-income households. Overall, the aggregate scale of these expanded unemployment benefits was similar to the Economic Impact Payments, with roughly $260 billion in expanded benefits paid out from April through the end of July 2020. However, unlike Economic Impact Payments, these $600 payments were targeted at households with declines in labor market income. These $600 supplements nearly tripled typical benefit levels and resulted in benefits that replaced about 145 percent of lost income for the median worker.

Ganong et al. (2021) use JPMCI data to show that expanded unemployment benefits substantially boosted the spending of unemployed households, while having a comparatively muted effect on job search at the time. They show that while the $600 supplements were available, the net income and spending of unemployed households actually rose rather than declined after job loss, both in absolute terms and relative to that of employed households. Using a variety of identification strategies, they estimate causal spending responses to the start of the $600 unemployment insurance supplement in April 2020, to its expiration in August, as well as to additional short-term supplements of $300 that were paid out in September. For example, they estimate causal responses to the start of expanded unemployment

Figure 4
The Timing of Unemployment Insurance and Increased Spending



Source: Ganong et al. (2021)

benefits by comparing the spending and income of households who all become unemployed at the end of March 2020 but begin receiving unemployment benefits at different dates.

The spending of these different cohorts is nearly identical up to the start of unemployment benefits and then jumps immediately when benefits begin. Exploiting these differences across cohorts, they estimate a marginal propensity to consume out of benefits of 0.43. Strikingly, they also find high marginal propensity to consume at the time of benefit expiration in August as well as in response to an additional short-lived $300 benefit increase paid out in September 2020, even though unemployed households had built up substantial liquidity by then through prior receipt of expanded benefits.

In contrast to these large spending responses, Ganong et al. (2021) find small effects of the $600 on job finding. Simple job finding models calibrated to pre-pandemic evidence predict a very large and sustained increase in job finding after the expiration of unemployment benefits—the job finding rate was quite stable from May through October. Moreover, they find an important role for recall to previous employers, rather than transitions to new employers, in explaining what fluctuations in job finding rates do exist over this period. Fitting a job search model with various elements to match the patterns in job finding and recalls, they estimate that employment distortions induced by the $600 unemployment insurance supplements were much lower than implied by pre-pandemic distortion estimates. They

also use this data to document a number of novel labor market facts that traditional data sources do not measure: for example, traditional unemployment data does not track individuals, but they show that repeat unemployment is particularly important during the pandemic.

Overall, this research leverages administrative account-level data in crucial ways, which cannot be done using other datasets currently available, by linking account-level income measures to account-level measures of spending and saving. However, this account-level data is ill-suited for answering certain other questions of great interest. For example, what fraction of households who lost jobs received unemployment insurance and how long do they have to wait to receive benefits? It might seem that bank account data could be used to answer this question, but in fact a large fraction of individuals now receive their unemployment benefits via pre-paid debit cards. These cards are unobserved in JPMCI data, and given their transitory nature, they are also unlikely to be linked in financial account aggregators like SaverLife. As a result, this financial account data cannot distinguish a worker who is waiting-for/denied/never-filed-for unemployment benefits from one is currently receiving unemployment benefits via a prepaid card.

Cell Phone Data

Cell phones produce near-constant streams of data that allow for detailed information on geographic location at very high temporal frequencies in near real-time. During the pandemic, movement and social interaction was of even more direct interest than usual. This data can also be used to proxy for shopping activity in narrow geographic areas, which can be used to identify the effects of government shutdown and reopening policies.

Couture et al. (2021) use data on roughly 75 million unique cell phones from PlaceIQ to construct a "daily location exposure index" that captures county-to-county movements together with a "device exposure index" that measures the exposure of devices to each other within venues. These exposure indices, which they post publicly every weekday, have been used in a variety of papers. In addition to providing this public good, Couture et al. (2021) provide an extensive discussion of representativeness and advantages and disadvantages of this cell phone data in addition to documenting several interesting patterns of movement during the pandemic.

I will not repeat that discussion here but will highlight a few key observations. First, the data is broadly representative of general population distributions and flows across space when benchmarked against various external data sources, but it is more representative when studying broader geographic areas like counties or states than when studying vary narrow geographies. Second, the PlaceIQ and most other US-based cell phone data is collected through smartphone apps with location-tracking services rather than directly from cell-service providers. Because older adults are less likely to own smartphones, older households are less represented in this data. This can, in turn, be important in studies of the pandemic, given that COVID-19 exhibits a sharp age-gradient in disease outcomes. They also caution

that cross-location level comparisons are likely to be less reliable than time-series variation within locations across time due to differences in coverage and representativeness across space. Furthermore, device IDs turn over frequently, which means that the panel element at the level of individual devices is typically limited to around six months.

Couture et al. (2021) is primarily focused on the development and validation of their exposure indices rather than on particular applications, but they do demonstrate a number of interesting results. During the pandemic, for example, the indices show that a sharp decline in movement in and out of Manhattan is detectable in near real-time in the early stages of the pandemic. More generally, they also explore the relationship between cell phone visit data and credit card spending data, which track each other very closely in some categories like arts and entertainment but sharply diverge in other categories like grocery spending. This divergence in grocery spending likely reflects a substitution towards online purchases together with the consolidation of multiple trips with smaller expenditure into single trips with larger expenditure per trip. Thus, while visits and spending generally track each other, this is not uniformly true. As a result, questions focused on physical presence, like in-person shopping time, are likely to be more reliably answered with cell phone data than questions about ultimate expenditures.

Goolsbee and Syverson (2021) use similar cell phone data from 45 million cell phone users produced by SafeGraph to try to understand the factors driving declines in consumer traffic from March 1 to May 16.[5] In particular, they seek to differentiate the role of government-imposed restrictions from households voluntarily staying home in driving changes in consumer behavior. To do this, they combine local store visit data from SafeGraph with county- and city-level shutdown policies and implement a cross-border identification strategy, which compares weekly shopping visits across counties with different restrictions within commuting zones. In particular, commuting-zone fixed effects should help to control for unobserved factors, like health fear, that are common to the commuting zone within that week. Thus, the effects of government restrictions will be identified only from variation in consumer behavior across counties with different policies all within the same commuting zone. This identification strategy reduces the concern that correlations between government restrictions and declines in consumer activity reflect a common response to rising health risk rather than a causal effect of the restrictions themselves.

Overall, they find that while consumer traffic fell by 60 percentage points, legal restrictions explain only 7 percentage points of this decline, which means that declining economic activity was predominantly driven by direct consumer responses to the virus rather than by government shutdown policies. Of course, this result needs to be interpreted with the usual caveat that cross-sectional causal effects may differ from aggregate effects; for example, any restrictions in one county that have

[5]Updates after publication of the original paper, available at (https://bfi.uchicago.edu/insight/research-update-drivers-of-economic-decline/), extend the analysis to the period of re-opening in the summer and to the second round of closures in the fall and find nearly identical results.

spillover effects on the commuting zone as a whole will be missed by this empirical strategy. Nevertheless, such spillovers would need to be implausibly large to undo the main message. Also, it is important to reiterate the point raised above that cell phone data captures store visits, not expenditures. However, Alexander and Karger (2020) use a similar identification strategy with store-level credit card spending data from Womply and arrive at very similar conclusions.

While the applications discussed in this section used a variety of different data sources, several lessons emerge. First, the pandemic caused substantial declines in spending which were very concentrated in certain service sectors. Second, these sectors employ many low-income households, so the pandemic led to much larger unemployment for low-income households. Third, these declines in spending and increases in unemployment were largely unavoidable in the sense that they were caused directly by health fears rather than by mandated government shutdowns. Fourth, US government transfers, in the form of broad-based stimulus checks and expanded unemployment benefit checks, led to substantial increases in spending for many of those otherwise hardest hit by the recession.

Additional Discussion and Conclusions

Some of the research results from using administrative micro data rely crucially on the particular lenses offered by this data. In other cases, similar results can eventually be obtained using traditional datasets, but the use of administrative data can allow research to proceed more quickly in a way that can offer more timely input for policy decisions. I close with a few additional reflections on the challenges facing researchers who are contemplating the use of administrative micro data.

First, administrative data can be difficult to interpret because samples are often unrepresentative and because they may contain a limited set of covariates of interest. For this reason, I stress that administrative micro data should be viewed as a complement rather than substitute for traditional data sources. Without representative data sources for benchmarking and validation, it is very difficult to interpret results from particular administrative data sources. Indeed, there has also been a dramatic expansion of traditional data surveys themselves during the pandemic. For example, the Census now conducts rapid household and business pulse surveys. Some of this data has, in turn, been used for similar near real-time analysis (as in Dube 2021).

Second, administrative datasets are often very large and thus will require substantial computational resources and time to analyze. Their size can also exacerbate the first concern about interpretation because it can be harder to notice data anomalies and other issues when the data itself is cumbersome to analyze.

Third, administrative datasets often have significant barriers to access: they may be expensive to purchase or might depend on personal connections for access. Further, continued access and data availability is often not assured. For example, data access can disappear because the provider goes out of business, changes

business models, becomes subject to new legal restrictions, changes licensing terms, or for many other unanticipated reasons. Even public-spirited firms quail at the prospect of making a commitment to maintaining these data-sharing arrangements over a period of years for each research project they approve.

These issues with access and the institutional risk of private-sector administrative data raise some concerns for the economics profession. Research directions may be overly influenced by the interests of a small number of individuals privileged with access. Other researchers may have limited or no ability to test the reproducibility and robustness of findings and to carry out extensions of the analysis. Journals are increasingly requiring detailed replication code and detailed data access instructions in online data repositories as conditions of publication. These repositories and associated access information are often a useful starting point for those interested in using the same data for follow-up work, but they do not themselves eliminate access barriers.

However, other kinds of administrative data are becoming more publicly available. For example, some academic institutions have been taking on a role as data intermediaries. The Kilts Center at the University of Chicago Booth School of Business acquires data from AC Nielsen and other private providers and then administers widely available academic licenses for this data. During the pandemic, Opportunity Insights began publicly publishing data that they obtain from a very large variety of private data providers (Chetty et al. 2020), although confidentiality agreements mean that this is not micro data and is instead aggregated to zip code or higher levels of aggregation. Finally, in response to the pandemic, many data providers have reduced the barriers to entry for acquiring micro data. For example, the SafeGraph data discussed above is now widely available for academic use.

Some government institutions have also embraced the use of administrative micro data and might play a role in this process. The Federal Reserve, the Consumer Financial Protection Bureau, the Office of the Comptroller of the Currency, and other regulators have access to a wide variety of administrative micro data, and they increasingly allow the use of this data to enrich their internal research: for example, Aladangady et al. (2019) discusses the potential of high-frequency administrative data for informing public statistics. These institutions have also increasingly provided paths for external researchers to access this data. By acting as an intermediary between private data providers and the research community as a whole, this model can potentially break down some barriers and democratize access to this data.

For individual researchers, working with administrative data often requires a significant investment of time and a degree of risk. Nevertheless, this comes with the opportunity for transformative research insights, which could not be made with other sources.

References

Aladangady, Aditya, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia Sahm. 2019. "From Transactions Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending." https://www.nber.org/books-and-chapters/big-data-21st-century-economic-statistics/transactions-data-economic-statistics-constructing-real-time-high-frequency-geographic-measures.

Alexander, Diane, and Ezra Karger. 2020. "Do Stay-at-Home Orders Cause People to Stay at Home? Effects of Stay-at-Home Orders on Consumer Behavior." https://doi.org/10.21033/wp-2020-12.

Baker, Scott R. 2018. "Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data." *Journal of Political Economy* 126 (4): 1504–57.

Baker, Scott R., R.A. Farrokhnia, Steffen Meyer, Michaela Pagel, and Constantine Yannelis. 2020. "Income, Liquidity, and the Consumption Response to the 2020 Economic Stimulus Payments." NBER Working Paper 27097.

Bartik, Alexander W., Marianne Bertrand, Feng Lin, Jesse Rothstein, and Matthew Unrath. 2020. "Measuring the Labor Market at the Onset of the Covid-19 Crisis." *Brookings Papers on Economic Activity* 1–72.

Bartlett, Robert P., Adair Morse, Richard Stanton, and Nancy Wallace. 2019. "Consumer Lending Discrimination in the FinTech Era." https://doi.org/10.2139/ssrn.3063448.

Beraja, Martin, Andreas Fuster, Erik Hurst, and Joseph Vavra. 2019. "Regional Heterogeneity and the Refinancing Channel of Monetary Policy." *Quarterly Journal of Economics* 134 (1): 109–83.

Berger, David, Konstantin Milbradt, Fabrice Tourre, and Joseph Vavra. Forthcoming. "Mortgage Prepayment and Path-Dependent Effects of Monetary Policy." *American Economic Review*.

Cajner, Tomaz, Leland D. Crane, Ryan A. Decker, John Grigsby, Adrian Hamins-Puertola, Erik Hurst, Christopher Kurz, and Ahu Yildirmaz. 2020. "The U.S. Labor Market during the Beginning of the Pandemic Recession." *Brookings Papers on Economic Activity* Summer (2020).

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Michael Stepner, and The Opportunity Insights Team. 2020. "How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data." NBER Working Paper 27431.

Cortes, Guido Matias, and Eliza C. Forsythe. 2020. "The Heterogeneous Labor Market Impacts of the Covid-19 Pandemic." Upjohn Institute Working Papers 20-327.

Couture, Victor, Jonathan I. Dingel, Allison Green, Jessie Handbury, and Kevin R. Williams. 2021. "JUE Insight: Measuring Movement and Social Contact with Smartphone Data: A Real-Time Application to COVID-19." *Journal of Urban Economics*.

Cox, Natalie, Peter Ganong, Pascal Noel, Joseph Vavra, Arlene Wong, Diana Farrell, Fiona Greig, and Erica Deadman. 2020. "Initial Impacts of the Pandemic on Consumer Behavior: Evidence from Linked Income, Spending, and Savings Data." *Brookings Papers on Economic Activity*: 35–82.

Farrell, Diana, and Fiona Greig. 2015. *Weathering Volatility: Big Data on the Financial Ups and Downs of U.S. Individuals*. JP Morgan Chase Institute. https://www.jpmorganchase.com/content/dam/jpmc/jpmorgan-chase-and-co/institute/pdf/54918-jpmc-institute-report-2015-aw5.pdf.

Dube, Arindrajit. 2021. "Aggregate Employment Effects of Unemployment Benefits during Deep Downturns: Evidence from the Expiration of the Federal Pandemic Unemployment Compensation." NBER Working Paper 28470.

Ganong, Peter, Fiona Greig, Max Liebeskind, Pascal Noel, Daniel M. Sullivan, and Joseph Vavra. 2021. "Spending and Job Search Impacts of Expanded Unemployment Benefits: Evidence from Administrative Micro Data." Becker Friedman Institute for Economics Working Paper 2021-19.

Ganong, Peter, Pascal Noel, and Joseph Vavra. 2020. "US Unemployment Insurance Replacement Rates during the Pandemic." *Journal of Public Economics* 191.

Goolsbee, Austan, and Chad Syverson. 2021. "Fear, Lockdown, and Diversion: Comparing Drivers of Pandemic Economic Decline 2020." *Journal of Public Economics* 193.

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

Grigsby, John, Erik Hurst, Ahu Yildirmaz and Yulia Zhestkova. 2021. "Nominal Wage Adjustments During the Pandemic Recession." *American Economic Review: Papers and Proceedings* 11: 258–62.

Johnson, David S., Jonathan A. Parker, and Nicholas S. Souleles. 2006. "Household Expenditure and the Income Tax Rebates of 2001." *American Economic Review* 96 (5): 1589–1610.

Mian, Atif, and Amir Sufi. 2018. "Finance and Business Cycles: The Credit-Driven Household Demand Channel." *Journal of Economic Perspectives* 32 (3): 31–58.

Parker, Jonathan A., Nicholas S. Souleles, David S. Johnson, and Robert McClelland. 2013. "Consumer Spending and the Economic Stimulus Payments of 2008." *American Economic Review* 103 (6): 2530–53.

Stroebel, Johannes, and Joseph Vavra. 2019. "House Prices, Local Demand and Retail Prices." *Journal of Political Economy* 127 (3).

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 67–82

Some Thoughts on the Washington Consensus and Subsequent Global Development Experience

Michael Spence

In 1989, policymakers around the world were struggling to come to grips with the debt crisis and slow growth that had plagued developing economies during much of the 1980s, especially nations in Latin America and sub-Saharan Africa. The International Institute of Economics (now the Peterson Institute of International Economics) held a conference discussing the economic and debt situation, mostly focused on Latin American countries. The conference was run by John Williamson (who died in April 2021), a senior fellow at the institute who specialized in topics related to international capital flows, exchange rates, and development. To focus the conference discussion, Williamson (1990) wrote a background paper that began: "No statement about how to deal with the debt crisis in Latin America would be complete without a call for the debtors to fulfill their part of the proposed bargain by 'setting their houses in order,' 'undertaking policy reforms,' or 'submitting to strong conditionality.' The question posed in this paper is what such phrases mean, and especially what they are generally interpreted as meaning in Washington."

Williamson (1990) described what he saw as a convergence of opinion about ten policies areas designed to promote stability and economic development that he felt had emerged during the 1980s. With hindsight, it appears that one of the principal targets was bouts of instability in inflation, public finances, and the balance of payments. If one asks who the consenting parties are in this "consensus," the answer appears to include the US Treasury, the International Monetary Fund and

■ *Michael Spence is Philip H. Knight Professor Emeritus, Graduate School of Business, and Senior Fellow, Hoover Institution, both at Stanford University, Stanford, California. His email address is mspence@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.67.

World Bank, think tanks with related agendas, to some extent academia, and over time Latin American governments who came to understand the destructive power of macroeconomic instability with respect to growth. It is noteworthy that in the mid-1990s, inflation in a wide range of developing countries dropped substantially and stayed there.

Williamson's original paper was organized around ten subject areas, but in a later essay, he usefully collapsed them into a list (Williamson 2004), which I reproduce here:

1. Budget deficits . . . should be small enough to be financed without recourse to the inflation tax.
2. Public expenditure should be redirected from politically sensitive areas that receive more resources than their economic return can justify . . . toward neglected fields with high economic returns and the potential to improve income distribution, such as primary education and health, and infrastructure.
3. Tax reform . . . so as to broaden the tax base and cut marginal tax rates.
4. Financial liberalization, involving an ultimate objective of market-determined interest rates.
5. A unified exchange rate at a level sufficiently competitive to induce a rapid growth in nontraditional exports.
6. Quantitative trade restrictions to be rapidly replaced by tariffs, which would be progressively reduced until a uniform low rate in the range of 10 to 20 percent was achieved.
7. Abolition of barriers impeding the entry of FDI (foreign direct investment).
8. Privatization of state enterprises.
9. Abolition of regulations that impede the entry of new firms or restrict competition.
10. The provision of secure property rights, especially to the informal sector.

This list was subsequently augmented by Dani Rodrik (2002) to include an additional ten areas of reform that are also correlated broadly with successful development and growth patterns in the post–World War II period, including issues of corporate governance, anti-corruption, flexible labor markets, and more. Both the Williamson and Rodrik versions, at some level of generality, make perfectly good sense. As such, if viewed directionally as a general guide to practitioners in thinking about reform agendas, the Washington Consensus seems relatively free of objectionable items. Many authors and commentators have expressed essentially this view.

However, the idea of a Washington consensus has also proven to be a flashpoint for controversy, which was typically less about the actual ten items than it was about the name of the list, what the list left out, and what implications to draw from the list.

Looking back, John Williamson regarded the word "Washington" in "Washington Consensus" as an unfortunate choice for many reasons. It suggested that

development policies were promulgated or mandated in Washington and hence externally imposed, though the consensus was meant to include policymakers in developing countries, certainly in Latin America, and, perhaps, more broadly. As Williamson (2004) wrote more than a decade later: "I labeled this the 'Washington Consensus,' sublimely oblivious to the thought that I might be coining either an oxymoron or a battle cry for ideological disputes for the next coup."

At a conceptual level, the Washington Consensus list was never intended to be interpreted as a fully elaborated plan, a growth strategy, or a model of development. A growth strategy is a complementary set of actions, reforms, and investments, with appropriate sequencing and pacing and is adapted to the specific initial (and partially historically determined) economic, social, and political conditions in a particular economy and society. In conjunction with a realistic model of how the economy will respond, a growth strategy will predict improved performance in terms of growth and economic development. One implication is that growth strategies are specific to particular countries and time: across countries, they may have common elements, but they must have idiosyncratic elements too. Williamson, an expert in development, knew all this. Rodrik, the leading development economist of his generation, was and is well aware that a few policy guidelines do not constitute a growth model nor a growth strategy. The Washington Consensus was never intended as a complete or a one-size-fits-all development program.

The ten-item policy list did not and does not purport to be a statement of either necessary or sufficient conditions for growth and development. Some might view them as quasi-necessary conditions, meaning if there is some significant deficit (lack of openness and connection to the global economy for example) in any one or a subset of them, then economic performance will be impaired. Several items on this list, however, prominently items 4 (financial liberalization) and 5 (a unified exchange rate), do not seem to be consistent with strategy and performance in a wide range of successful developing countries, especially those in Asia.

It would be even more problematic to view the list as a set of sufficient conditions. Political and policy leaders in developing countries, many researchers, and academics understand that we don't know the sufficient conditions for growth: that is, we do not now (and did not then) possess models that fully capture the complex economic and political economy dynamics associated with growth and development. This may sound esoteric, but it is important. It means that even if a country does all ten items on the Washington Consensus list, there is no guarantee that growth will accelerate. Conversely, there may be and probably are multiple growth strategies that work reasonably well.

One of the persistent problems with development policy discussions is the absence of an explicit accompanying growth model. The protagonists in development debates often appeal to their own models, or just leave it vague. The Washington Consensus has been misused, both by those appealing to its authority and those rebelling against it, in the service of their own preferred growth models. In this essay, I look back at the Washington Consensus in terms of what we have learned and experienced about economic growth since the late 1980s. I also seek to

stipulate what the Washington Consensus was, was not, and what (as far as I know based on John Williamson's writings) it was never intended to be. In the other three papers in this symposium, Anusha Chari, Peter Blair Henry, and Hector Reyes test the hypothesis that countries which enacted Washington Consensus reforms tended to experience faster growth in the following decade; Ilan Goldfajn, Lorenza Martínez, and Rodrigo Valdés discuss the implementation and legacy of the Washington Consensus reforms in Latin American countries; and Belinda Archibong, Brahima Coulibaly, and Ngozi Okonjo-Iweala consider the implementation and legacy of Washington Consensus reforms in countries of sub-Saharan Africa.

In this paper, I want to view the Washington Consensus through the lens of subsequent growth and development experience as well as related research across the developing world. The trajectories of a number of Asian economies, before and after the Washington Consensus was written, provide useful lessons. At the end of the paper, I will return to uses and misuses of the Washington Consensus, and, specifically, a slimmed down version of the Consensus that was used to justify writing government out of too many aspects of the development strategy script.

The Washington Consensus and Growth 15 Years Later

My views on the relationship between the Washington Consensus and growth policy were shaped by my experience from 2006 to 2008 in chairing the Commission on Growth and Development (2008).[1] Its mandate and purpose were to review development progress on a global basis. After all, a number of development stories had emerged around the world in the roughly decade and a half since the Washington Consensus was formulated. China had sustained average growth at or above 8 percent for 25 years. India's growth experienced a notable acceleration starting in 1991. Brazil had gone from high growth in the two decades after World War II to two decades of economic and political turmoil in the 1970s and 1980s, but had overcome destructive hyperinflation and seemed to be in the process of restoring growth momentum. In east Asia, South Korea and Taiwan had engineered largely successful transitions from middle-income to high-income status, notwithstanding the negative shock of the 1997–1998 Asian financial crisis, and Vietnam had experienced accelerating growth and integration into the global economy. We sought to assess what had been learned from experience in a wide range of countries (those that had experienced rapid growth and poverty reduction and many that had not) and from academic and policy research.

Many of the observations that emerged from that exercise are consistent with the Washington Consensus viewed as policy guidelines, provided one recognizes that their relative importance fluctuates with the variations in the context of a specific

[1] As I note in my "Preface" to the report: "This report brings together the views of a Commission of 19 leaders, mostly from developing countries, and 2 academics, Bob Solow and me" (Commission on Growth and Development 2008). We also received support from World Bank staff.

country's conditions. But the emphasis of our main growth lessons was somewhat different.

In search of common elements of sustained growth experiences, we found six key areas. It may be useful to state them here as context for the more detailed remarks below. They were macroeconomic stability, exploitation of world markets and technology/knowledge, high levels of investment and saving, allowing markets to play a role in resource allocation and incentives, leadership and governance, and finally, managing the distributional aspects of growth patterns to put boundaries on inequality in various dimensions.

A first observation was that the demand in the global economy, and specifically, its enormous size relative to any early stage developing economy, is crucial. Domestic demand in a low-income country, both its size and composition, is a severe constraint to sustained productivity growth that enables overall growth. In isolation, absent international trade and the resulting specialization, domestic demand and supply have to coincide. It would be as if the entire economy was located in the non-tradable sector. Drivers of productivity like scale economies, learning curves, and even exploiting imported technology are all truncated in the non-tradable sector of a low-income economy. Domestic demand does not support specialization. As far as I know, there are no cases of sustained relatively high growth that are not export- and trade-enabled. The Washington Consensus on opening via lowering trade restrictions both in and outbound (item 6 on reducing trade barriers and item 7 on foreign direct investment) aligns with this reality.

However, as many authors have noted, integrating with the global economy does not mean sudden shifts in patterns of openness are appropriate. Rapid shifts may occur too quickly for the domestic economy to adjust structurally, creating economically and socially damaging disruption and unemployment, which in turn, may undermine the political support for reform agendas. In this and other reform areas, both pace and sequencing are important. Economic theory is not particularly helpful in this area, because most economic theory deals with equilibria, not transitions between equilibria. This means that for policymakers, economic theory is more helpful in determining the destination and where you want to go, and less helpful as a guide as to how exactly to get there. Pace and sequencing are more a matter of judgment and art than science. If political economy is partly about feedback loops amongst economic policy, economic outcomes, and political/electoral outcomes, then the Washington Consensus can be seen as essentially devoid of political economy considerations.

A second insight from studying and observing growth cases is that the global economy, particularly more advanced countries, provided technology that when absorbed and adapted in a developing economy, causes potential productivity and output to grow much more rapidly than it would or could if the technology had to be generated endogenously from within.

Knowledge transfer is an incredibly powerful accelerator of potential growth. Paul Romer's (1994) work on endogenous growth explains why this is true and how

it works.[2] Endogenous and self-generated technological advance occurs in advanced economies and underpins their growth. But in the early stages of growth in developing countries, self-generated technology is heavily supplemented by inbound technology transfer, enabled by the technological divergence between advanced and developing countries that grew over two centuries since the Industrial Revolution. For this reason, developing-economy growth is frequently referred to as "catchup growth." More than any other factor, this explains why we see sustained growth rates in excess of 7 percent in some developing countries. The Washington Consensus is almost, if not quite completely, silent on this. Inbound foreign direct investment (in item 7) often is (or can be) an important channel for the inbound transfer of practical knowledge, technology, and know-how. In my view, a major weakness of the Washington Consensus as a guide to development policy formation is an under-emphasis on knowledge transfer and the channels through which it occurs, as well as on the domestic conditions and investments that facilitate absorption and diffusion of knowledge and technology. Indeed, the country-specific adoption and application of some subset of the common elements of successful growth strategies can be thought of as its own form of knowledge transfer among developing countries, creating an environment in which companies and/or governments can usefully import and embrace the new production technologies embodied in foreign direct investment and other channels.

A third theme from the Commission on Economic Development (2008) report is the very close connection between structural transformation and economic growth. The Washington Consensus has a rather pronounced macroeconomic focus, probably because it was informed by the numerous instances of high debt and destructive macroeconomic instability at the time. But this emphasis is still somewhat puzzling. Sir W. Arthur Lewis (1954) set forth a sectoral-based developing country growth model that was widely known at the time, and for which he had received the 1979 Nobel Prize in economics ten years earlier.[3] At the core of the Lewis framework is structural change in the economy. Specifically, in early stages, growth is driven by productivity growth, and hence, income growth in the expanding tradable sectors via exports, drawing labor from agriculture and related traditional sectors. This structural change is not a side-effect of growth, but the key element in the growth dynamics.

One can only speculate about the lack of specific reference to structural change in the Washington Consensus. Perhaps at the time, a strong view in Latin America and parts of academia—in the context of the debt crisis and high inflation rates—was that markets by themselves in a properly regulated and relatively stable macro environment would take care of structural change. But this belief is not written into the Washington Consensus; instead, it is part of a growth model that can be strongly

[2]For background on endogenous growth theory, Aghion and Howitt (1992) is a useful starting point. Also, the Winter 1994 issue of this journal includes a four-paper symposium on "New Growth Theory." Along with the contribution from Paul Romer (1994), it includes essays by Gene Grossman and Elhanan Helpman, by Robert Solow, and by Howard Pack.

[3]For a 60-year retrospective on the Lewis model in this journal, see Gollin (2014).

disputed. To be sure, private sector incentives, investment, and dynamics are important elements in structural transformation and growth. But they are not the whole story. Allocating structural change entirely to the private sector seems to miss or ignore the role that both the size and composition of public sector investment in human capital, infrastructure, technology, public goods, urbanization, migration policies, and social security systems in general play in affecting the size and direction of structural change.

In particular, public sector investment is an essential element of growth and development dynamics. The main elements are human capital, infrastructure, and the knowledge and technology base of the economy. These investments have high social rates of return precisely because and when they raise the rates of return to private investment (both domestic and foreign) in the private sector. In the Washington Consensus list, this would correspond to item 2, which contains an added and important twist. It says that governments should stop spending limited public resources on inefficient and wasteful subsidies and devote them to productivity-enhancing social investments. The public sector also plays a vital role in protecting people from the most adverse outcomes that go along with rapid structural change. The presence or absence of such policies will feed back, positively or negatively, on public support for the overall growth-oriented reform agenda.

This brings us to a fourth and more general point about the inclusiveness of growth patterns. There is one point, essentially missing in the Washington Consensus, on which there was unanimous agreement among the policymakers from around the world who were members of Commission on Economic Development (2008): Non-inclusive growth patterns generally fail. Put differently, growth that is accompanied by extreme hardship for large groups due to the turbulence of creative destruction that accompanies structural change (think of large-scale loss of employment), by rapid increases in inequality, or by cases of large-scale inequality of opportunity or access, will encounter resistance, and eventually, the likelihood that the policy underpinnings will be rejected rises. Therefore, the public sector plays a critical role in the design and implementation of reform programs, with an eye to preventing excessively non-inclusive outcomes.

This omission is somewhat puzzling, at least to me. To the extent that the ideas embodied in the Washington Consensus were informed primarily by experience in Latin America, where some of the highest national levels of income inequality could be found, and where the political economy saw political polarization and wide ideological swings from populism to market fundamentalism, one might have expected that the inclusiveness of the growth patterns, or its absence, might have made its way to the policy guidelines.

Sometimes the structural changes produced by market outcomes are relatively benign, as in the case of immediate post–World War II growth in developed countries (say, from 1945 to 1970). But more recent history instructs that this is not always the case. An important part of the role of the state, as a complement to otherwise beneficial market forces and incentives, is to engineer and nudge the growth trajectory in order to contain inequality and exclusion and to promote intergenerational

mobility. Among the instruments are universal delivery of key public services, especially education and health, but also access to financial services and a broad-based method of taxation. Connectivity via physical and information technology infrastructure, an important element of public sector investment, can also have beneficial distributional benefits if properly implemented. Of course, these things are expensive and cannot be done on short time horizons. But attaching a priority to making discernible progress on them does enhance overall growth while mitigating inequality.

In fairness, it should be noted that just as a country can have too much emphasis on growth as well as static and dynamic efficiency at the potential cost of adverse distributional trends and eventual opposition to the growth agenda, a country can also have too much focus on distribution, and too little on growth itself and the contribution of the private sector's key role in structural change and advancing productivity. Some of the Washington Consensus looks like it is meant to lean against this second tendency (as in items 5–9). After all, growth is a necessary condition for rising incomes, opportunity, and poverty reduction in lower-income countries. One of the main shortcomings of populist governments, at least in some of their manifestations, is that they leverage public sentiment around distributional problems while either ignoring the longer-term growth agenda, or worse, taking policy actions that adversely affect growth.

A common, indeed nearly universal, feature of development policies at the time, and even later, were subsidies especially for fossil fuels and sometimes electricity. These are counterproductive from the point of view of dynamic and static efficiency; in addition, viewed as a negative tax, they are probably regressive and by distorting the price signals, they guide the economy to low energy-efficiency and high energy-consumption paths, which affect the patterns of long-lived capital investment. In 1989, climate change was not widely perceived as the existential global challenge that it has become now. Subsidies to fossil fuels, in retrospect, seem even worse than they did then.[4] When I had the chance to talk with political leaders, they understood that subsidies were counterproductive, but they also knew that such subsidies are politically very difficult to remove once in place. Also, they are frequently implemented by governments via price controls on domestic energy products, an approach that hides the fact that the government is, in effect, giving up tax revenue that could have been spent more productively.

The Washington Consensus and Asian Development Experience

The Washington Consensus reform agenda, which I prefer to think of as a set of guidelines for reasons explained earlier, seems to have been informed mainly by

[4] In the Commission on Economic Development (2008) report, perhaps the most popular section was a two-page discussion of "Bad Ideas" (pp. 68–69). The first bad idea listed was "[s]ubsidizing energy except for very limited subsidies targeted at highly vulnerable sections of the population."

experience in Latin America, and specifically by addressing bouts of fiscal and financial instability. However, countries in Asia have, on average, outperformed the rest of the developing world by a fairly large margin in terms of sustained growth over the last few decades. Back in 1989, while China had entered an economic reform phase ten years earlier, it was not yet clear as it is now that a decades-long period of unprecedented growth had been launched. The economies of South Korea and Taiwan had performed well, but in the 1980s, they were in the midst of the perilous middle-income transition. It was not at all that clear then that they would sustain growth to achieve developed economy income levels.

The Washington Consensus, as far as it goes, is broadly consistent with Asian development strategies. However, several items on the list—like item 5 on exchange rates and capital account management and item 8 on privatizing state-owned enterprises and generally getting the government out of specific sectors—do not seem in accord with all or most Asian policy choices. For me, juxtaposing the Washington Consensus development policy guidelines with experiences in a range of Asian countries/economies, before and after 1989, offers a way to think about what is not included in the Washington Consensus, and what development experience has taught us in the intervening 30 years.

Let's begin with economic theory and conceptual frameworks. Economic models were regarded as useful by policymakers in Asia. China's original request to the World Bank in the 1980s was for help in importing western knowledge about the management of a market economy.[5] But in China and other Asian economies, the models that are used in developed market economies to predict the outcomes of policy choices need to be handled with caution. The reason is that these models assume, mostly implicitly, a fairly fully developed set of market institutions and capacities. In the early stages of growth, these do not exist in fully developed form. Policymakers in China, for example, explicitly viewed the economy as a transitional one (and still do), where the transitions are multi-dimensional: structural, human capital deepening, building market and institutional depth and development (especially in finance), and more.

When beginning from this transitional mindset, the analytical tools of economics for predicting the impact of policies are not fully developed and the approach becomes what might be called pragmatically experimental. I have referred to this approach as akin to navigating with incomplete charts (Spence 2010)—not quite like the case of the early global maritime explorers who had no charts whatsoever, but incomplete in important ways.

Asian development policies generally were informed by explicit (and evolving) views about the sources of comparative advantage, and hence about what kinds of investment (public and private, foreign and domestic) were likely to be needed to access them. Asian economies (with some exceptions) generally are not rich in

[5]For an interesting discussion of what China hoped to get from its interactions with the World Bank and the interactions between Deng Xiaoping and Robert McNamara, see Edwin Lim's (2002, starting on p. 18) interview with the World Bank Oral History Program.

natural resources. However, these countries had an abundance of workers with relatively low incomes, and thus labor costs, and they had surplus labor in traditional sectors like agriculture. Thus, labor-intensive, process-oriented manufacturing and assembly (usually with textiles and apparel as the starting point) emerged as a key component of the growth model—the part that leveraged the global economy and specialization in the tradable sectors of the economy. At some level, these countries understood or came to understand the growth dynamics embedded in the Lewis (1954) growth model. In China, policymakers and their academic advisers talk explicitly about the "Lewis turning point" (as discussed, for example, in Das and N'Diaye 2013; Fang 2021), the point at which the shift of labor from traditional to modern urbanized sectors reaches a point that incomes and prices start to rise.

The more general point is that structural transformation and supporting policies are a central feature of development strategy discussions. Development policy in Asian countries tended to take a more expansive and flexible view of the role of government than is perhaps implicit in the ten-item Washington Consensus list. Government influence in Asian development experience included long time horizons, implemented via rolling five-year plans, which are best thought of not as plans but priorities for policy and development and statements about the direction of the economy. The goal was to solve coordination problems via providing a mechanism that helped expectations to converge. In France, this element of policy has been called "indicative planning." There was, in addition, a willingness to have government participate as a catalyst to structural change and growth at a microeconomic level, including via still-controversial industrial policies. The directions of public investment nudged the economy forward in terms of structural change. Most of Asia's policymakers appear to have understood the difference between crowding in and crowding out in their use of public sector investment. They learned over time the importance of the relation between expectations and coordination of economic activity. They also knew that internal and external shocks are to be expected, so that foreign exchange reserves, relatively modest government debt, and in some cases like China, even substantial state ownership of productive assets, came to be viewed as important tools in buffering shocks.

This general framework and the interventions that emerged from it were far from error-free. Mistakes are an automatic correlate of using judgment in the face of uncertainty. Good policy does not mean that mistakes never occur, only that they be promptly reversed. Generally, the formulation of policy was pragmatic and experimental, exhibiting less concern for any particular orthodoxy, and more concern for measurable progress toward explicit economic and social development goals. This distinction is quite fundamental. The goals and the system for achieving them are distinct. In all successful cases of high growth development, not just in Asia, markets, prices, incentives, decentralization, and capitalist investment and dynamics have been key elements of the system. But markets and free market policies should not be and were not confused with the objectives of development. They are instruments or tools. This mindset is particularly important when the mapping from policies to outcomes is highly uncertain. Asian policymakers generally maintained a focus on

the goals and had a flexible attitude toward which policies and tools would work and in what circumstances.

One implication of having a more expansive view of the role of government as a complement to a developing private sector was the need to have talent in both sectors, especially the public sector. Compensation and prestige both played a role in attracting it. Another is the need to stamp out corruption in its various forms. Nothing short-circuits development faster than a government that is either incompetent, or worse, pursuing some agenda and set of interests that diverge from the long-run public interest.

There are other differences between the Asian development experience and at least some interpretations of the Washington Consensus, though in reality they are not, for the most part, inconsistent with each other. The Washington Consensus approach to opening the economy in trade (item 6) and in the capital account (implied by items 4, 5, and 7) was somewhat cautious in Asia. On the trade side, my view is that in a highly uncertain transitional setting, Asian policymakers were trying to make sure the pace of opening was consistent with the capacity of the economy to adapt structurally, and specifically with a focus on the dynamics of the labor market and the balance between employment creation and destruction. Similar considerations apply to the capital account and exchange rates, which remain, to this day, a controversial area. There was and still is considerable variety across Asian economies in the management of the capital account. But there are few examples, if any, of totally open capital accounts and purely market-determined exchange rates. Inward foreign direct investment was generally favored as supportive of the growth model, though even here there are counterexamples; in one prominent case, Japan was not receptive to inward foreign direct investment for a number of decades and found other ways of accessing global technology. South Korea had a similar approach. Relatively more mobile (and potentially volatile) international financial flows are not without benefits, but they generally faced more restrictions. These restrictions tended to decline over time as the depth and liquidity of the capital markets increased and the capacity grew to absorb rapid shifts in these flows without risking instability.

Two points seem to me to emerge from these observations. First, if the Washington Consensus were to be rewritten or replicated in roughly the same time frame as the Tokyo Consensus, or perhaps later as the Shanghai Consensus, it would have looked similar up to a point, but there would be differences. It would have included more explicit recognition of, and emphasis on, the potential sources of comparative advantage and on the role of government in exploiting them. It would have been more explicit about the core features of the underlying growth model, the importance of knowing what those core features are in setting reform priorities, the evolving role of the state in catalyzing and facilitating structural change, the importance of policies targeted at inclusiveness in the growth process, and probably an explicit recognition that development is a multi-decade journey with extreme uncertainty at every step along the way.

Second, we are now living in a period of radical digital transformation of economies globally and of the global economy itself. This transformation has

many dimensions, but a few stand out as especially relevant for development. The expanding scope and scale of digitally enabled automation, powered by breakthroughs in machine learning, sensors, and more, means that robotics will sequentially overtake labor-intensive processes in manufacturing, logistics, and some related service functions in terms of cost. Digital technologies applied to automation and many other areas have high fixed cost and low-to-negligible variable or marginal cost. Thus, as scale increases, the average costs keep coming down and eventually take out and displace labor-intensive technologies.

This trend is well underway and is irreversible. It has profound implications for the location of manufacturing and the configuration of global supply chains. For developing countries, it means that the comparative advantage in labor-intensive manufacturing (the core of the "Asian development model") will decline and steadily lose its power as a growth engine. Rodrik (2015) refers to this process as "premature deindustrialization"—premature in the sense that it is a development path rapidly becoming unavailable. He has observed that manufacturing in a number of currently low-income countries is excessively capital intensive in the sense that capital-intensive or digitally capital-intensive manufacturing does not generate enough employment to support the demand and income side of the growth model. Early-stage developing countries will need to search for alternative sources of comparative advantage, ones that have powerful embedded employment engines. At present, the possible alternatives are not at all clear.

For middle-income countries, this digital trend may, on balance, be beneficial. A middle-income economy is already in a transition that involves moving people with higher levels of education away from the low labor-cost sectors or components of value chains to higher value-added activities, many in the growing service sectors or to service parts of value-added chains. In such countries, the automated parts of manufacturing may remain domestic, albeit with a much less labor-intensive configuration. With suitable human capital investment, the employment engines will shift to other parts of the economy.

Global trade in goods is in a period of decline, measured as a fraction of global GDP. But trade in services, although it is only about one-third of the trade in manufacturing/goods, is growing rapidly in absolute terms and as a fraction of GDP. There are valuable pools of relatively immobile human resources in a wide range of countries in the non-automatable parts of the global service sector. The global economy will find these workers and integrate them into global supply chains, unless we have a new bout of protectionism with rising barriers. The challenge for the lower-income countries is to find niches in this global services trade and adapt the policies and public sector investments to enter them. Development strategies will have to change.

There is a somewhat different set of digital technologies that show considerable potential with respect to inclusive growth. Research at the Luohan Academy (2019) in Hangzhou in China, using e-commerce and mobile payments data, indicates that platform-centered and open digital ecosystems can exhibit relatively powerful inclusive growth characteristics. For example, remote regions and lower-tier cities

gain access to markets and retail options that are not yet available in the offline world. With low entry barriers and supporting resources available via the platforms, entrepreneurial activity expands. Taobao, one of the principal e-commerce platforms in Alibaba, has 10 million companies and entrepreneurs on the platform, roughly 50 percent of whom are women. Vast troves of data in the e-commerce and mobile payments systems, when subjected to machine learning algorithms, are expanding credit to lower-income households and small businesses that were previously excluded from traditional credit channels because of lack of collateral and previously limited accessible financial histories. In economic terms, big data is closing informational gaps in some markets, with beneficial effects on market formation and efficiency.

These technological and market trends are not unique to China, though the digital infrastructure at this point is somewhat more advanced there than in lower-income countries. India, for example, is rapidly building similar digital economy systems around the rapidly growing Jio mobile phone network and expanding e-commerce platforms. E-commerce, mobile payments, and fintech platforms are expanding rapidly in Latin America as well. Africa has seen the development of innovative digital payments platforms. In short, digitally enabled or enhanced markets, commerce, and finance show considerable potential for becoming inclusive growth engines in developing countries and emerging economies.

Some Closing Thoughts

There is little controversy that the choice of the "Washington Consensus" as a name was unfortunate. It created a shadow with an unintended, vaguely imperialist connotation to what is otherwise an entirely thoughtful and insightful set of guidelines for thinking about development strategy and policy. The name made it a convenient target. If John Williamson (1990) had used some long-winded title like "Some lessons learned from experience in overly indebted developing countries, with special focus on disruptive bouts of instability caused by failures in macroeconomic management, and policies that help avoid them," then his list of ten policy guidelines probably would not have experienced such ferocious attacks. After all, there is no doubt that widespread financial crises before and during 1980s, and since then as well, have been a major impediment to progress in development in a wide range of countries. There is virtually no controversy about the importance of macroeconomic stability and the avoidance of self-inflicted wounds in the form of internally generated economic crises. They just slow down growth and development, and the recovery period is often lengthy. It is interesting to speculate whether the Asian economic and financial crisis of 1997–1998 might have been averted or been less severe if some of the Washington Consensus guidance had been heeded.

But the real heat directed toward the Washington Consensus came from a different direction. Somehow the Washington Consensus got linked with

development strategies, mainly in Latin America, that relied heavily on markets and private enterprise to generate growth and largely wrote government out of the script. Moreover, these development strategies paid little attention to issues related to pacing and sequencing of reforms and the shocks that might occur as a result, and that to a large extent ignored the distributional aspects of the growth patterns that might result. As I noted earlier, this last point, ignoring the distributional consequences of growth policies, is especially puzzling in a continent that had (and still has) some of the highest levels of inequality (for example, as measured by Gini coefficients) in the world. But on reflection, perhaps it is not that strange. After all, a multi-decade pattern of rising inequality in developed countries, especially the United States and United Kingdom, went largely unattended to in terms of policy countermeasures, at least until recently.

Historians will have to sort out how this linkage of the Washington Consensus with the limited government approach to development happened. As far as I can tell, it is essentially impossible to link the Washington Consensus as it is actually written with what has come to be called the neoliberal approach to growth and development. In particular, item 2 identifies a key role of government as an investor in infrastructure and human capital, a role for government well beyond the basic tasks like rule of law, defense, and sound macroeconomic management. But it is fair to point out that because the Washington Consensus was focused mainly on macroeconomic policy and stability, and to some extent on dysfunctional things that governments do that they should stop doing, it is largely silent on what is being assumed about the underlying growth model, what are its moving parts, and what roles do various sectors (foreign and domestic) play in getting the job done at various stages of development. Because of this, it perhaps became unintentionally vulnerable to the criticism that it had implicitly allocated most of the growth dynamics to the private sector. In addition, it seems clear that proponents of the neoliberal versions of the model with a limited role for government, on the ground that government is usually incompetent, wasteful, corrupt or all of the above, often viewed themselves as justified by the Washington Consensus policies.

My opinion is that if the Washington Consensus had been preceded by a preamble in which the key elements of a development model were laid out in such a way that the policy recommendations could be seen as implementing or partially implementing a growth strategy, it might have been interpreted differently: for example, the importance of leveraging global demand and technology could have been linked to items 6 (reduced trade barriers) and 7 (allowing foreign direct investment), the crowding in effect of properly targeted public investment could have been linked to item 2 (retargeting of government spending), and something on the high levels of public and private investment required to sustain elevated growth. The various pieces would have been seen as complementary components of an overall strategy. The guidelines would have been less susceptible to being treated as an à la carte menu, picking and choosing the items that conform to one's ideological predispositions and disposing of the rest.

The Washington Consensus was clearly well-intentioned, and in many ways insightful and a useful response to the accumulated experience at the time. With the benefit of hindsight, it was vulnerable to ideologically motivated misuse. That said, it has weathered the test of time pretty well. Subsequent experience and learning have not invalidated what it says in any major way, but instead have called attention to what it does not say and to some of the items that were often not followed. To be sure, there are subtleties and details that cannot be incorporated in a general set of guidelines because they are case-specific and to some extent idiosyncratic. The concept that policymakers in a given developing countries should seek to identify and address the binding constraints that apply to their own economy, as developed by Hausmann, Rodrik, and Velasco (2005), is a useful way of helping policymakers think about setting priorities in a specific time and place. These binding constraints are emphatically not the same as one moves from case to case, or even over time: for example, in a given country the key constraint could be demand shortfalls, deficits in human capital, or infrastructure.

The Washington Consensus has sometimes been criticized as promoting a one-size-fits-all approach to development. That complaint is unfair, and it was not Williamson's intention. Any attempt to distill lessons from experience across a range of developing economies, and even continents, would be vulnerable to the same objection. The truth is that successful development strategies and supporting policies are always context-specific. However, the fact that growth strategies cannot simply be written down or summarized in a list and transplanted in total from one setting to another does not mean that there are no common elements in successful development cases, nor does it mean that there is no value in cross-border learning. In fact, one of the more encouraging developments in the decades since the Washington Consensus has been the breaking down of regional silos within and between international financial institutions and the regional development banks around the world.

Although the world has lost the wise counsel of John Williamson with his death earlier this year, I am confident that the Washington Consensus, notwithstanding the controversy that has sometimes surrounded it, will come to be seen as an important milestone on a long and, at times, bumpy journey during which the welfare and the opportunities of hundreds of millions of people in the developing world have been lifted. In the dark days of addressing the immediate threat of a pandemic, it is well to remember both that much has been accomplished, and that there is much more to do.

■ *I would like to warmly thank the JEP editors, Heidi Williams, Erik Hurst, and Timothy Taylor, and also Peter Henry for giving me the chance to participate in this symposium and for a large number of thoughtful and useful comments and suggestions on earlier drafts. Any remaining deficiencies are solely the responsibility of the author. I also want to record that during the work of the Commission on Growth and Development, I was privileged to have a visit with John Williamson. He was gracious, supportive, and insightful. We shared the view that interim progress reports are fine, but that the learning process is continuous and does not have an end.*

References

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60 (2): 323–51.

Commission on Growth and Development. 2008. *The Growth Report: Strategies for Sustained Growth and Inclusive Development* Washington, DC: World Bank.

Das, Mitali, and Papa N'Diaye. 2013. "The End of Cheap Labor." *Finance & Development* 50 (2): 34–37.

Fang, Cai. 2021. *Understanding China's Economy: The Turning Point and Transformational Path of a Big Country*. Singapore: Springer.

Gollin, Douglas. 2014. "The Lewis Model: A 60-Year Retrospective." *Journal of Economic Perspectives* 28 (3): 71–88.

Hausmann, Ricardo, Dani Rodrik, and Andrés Velasco. 2005. "Growth Diagnostics." https://drodrik.scholar.harvard.edu/publications/growth-diagnostics.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labour." *Manchester School* 22 (2): 139–91.

Lim, Edwin R. 2002. "Transcript of Interview with Edwin R. Lim." http://documents1.worldbank.org/curated/en/344281468157786211/pdf/790750TRN0Lim000October030031002002.pdf.

Luohan Academy. 2019. *Digital Technology and Inclusive Growth*. Hangzhou, China: Luohan Academy.

Rodrik, Dani. 2002. "After Neoliberalism, What?" Paper presented at the Alternatives to Neoliberalism Conference sponsored by the New Rules for Global Finance Coalition, May 23–24.

Rodrik, Dani. 2015. "Premature Deindustrialization." *Journal of Economic Growth* 21:1–33.

Romer, Paul M. 1994. "The Origins of Endogenous Growth." *Journal of Economic Perspectives* 8 (1): 3–22.

Spence, Michael. 2010. *The Next Convergence: The Future of Economic Growth in a Multispeed World*. London: Picador.

Williamson, John. 1990. "What Washington Means by Policy Reform." In *Latin American Adjustment: How Much Has Happened?*, edited by John Williamson. Washington, DC: Peterson Institute for International Economics.

Williamson, John. 2004. "The Strange History of the Washington Consensus." *Journal of Post Keynesian Economics* 27 (2): 195–206.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 83–108

The Baker Hypothesis: Stabilization, Structural Reforms, and Economic Growth

Anusha Chari, Peter Blair Henry, and Hector Reyes

During the late 1970s and early 1980s, developing country governments from Kingston to Kuala Lumpur ran large fiscal deficits, causing their countries' stock of public debt to increase faster than GDP. As debt-to-GDP ratios breached critical thresholds and real interest rates for borrowing in US dollars rose, access to foreign financing ceased. When Mexico defaulted on its external obligations in 1982, precipitating a global debt crisis, governments increasingly turned to monetization as an alternative source of funding and inflation rose. By 1985, the average rate of inflation in the developing world was approaching 40 percent per year, with some countries spiking into hyperinflation.

On October 8, 1985, then-Secretary of the US Treasury, James A. Baker III, acting on a body of accumulated but untested knowledge about the potential benefits of economic policy reform whose origins lay with Krueger (1974), Balassa (1977), and others, unveiled a "Program for Sustained Growth" at the meetings of the International Monetary Fund (IMF) and World Bank in Seoul, South Korea. Baker (1985, p 207) said: "If the debt problem is going to be solved, there must be . . . First and foremost, the adoption by principal debtor countries of comprehensive

■ *Anusha Chari is Professor of Economics and Finance, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and Research Fellow at the Center for Economic Policy Research, London, United Kingdom. Peter Blair Henry is William R. Berkley Professor of Economics and Finance, Leonard N. Stern School of Business, New York University, New York City, New York. Hector Reyes is a Fellow, PhD Excellence Initiative, Leonard N. Stern School of Business, New York University, New York City, New York. Their email addresses are achari@unc.edu, pbhenry@nyu.edu, and hr133@stern.nyu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.83.

macroeconomic and structural policies, supported by the international institutions, to promote growth and balance of payments adjustment, and to reduce inflation." He enumerated a list of economic reforms—inflation stabilization, trade liberalization, greater openness to foreign investment, and privatization—that he urged "Third World" leaders to adopt, both as a way to reestablish their ability to borrow in international markets and also to enable their countries to grow again. These policies were later codified and branded "the Washington Consensus" by Williamson (1990).

Baker's speech unleashed a contentious and still unresolved debate about the economic impact of his recommended reforms. Opponents argue that the Washington Consensus failed (Rodrik 2006), and it has been disparagingly labeled as "neoliberalism" by others (for example, Chomsky 1999; Stiglitz 2002). Proponents contend either that reforms have been found difficult and left untried (Krueger 2004; Gil Diaz 2003), or that the results have been positive if comparatively modest (Easterly 2019; Grier and Grier 2021).

The persistence of this dispute is puzzling, given that Baker's speech would seem to constitute a testable claim that can be confronted with data: "If developing countries implement this set of reforms, then their standards of living will rise at a faster rate than they did before the implementation." The most common approach to evaluating this claim, however, has involved regressions with long-run growth rates (often measured by 30-year averages) as the dependent variable, and a dummy variable that indicates the presence or absence of certain policy reforms during the entire period over which growth is measured as the key explanatory variable (for example, openness to trade as in Rodríguez and Rodrik 2000). Cross-sectional regressions, however, provide a weak test of the hypothesis in question. Regressing countries' average long-run growth rates on policy-related dummy variables that are either "on" or "off" asks the following question: Is it the case that countries with low inflation, free trade, and liberalized capital accounts have higher long-run growth rates than countries with high inflation, restricted trade, and closed capital accounts?

The problem with this question is that a Solow-style (1956) model does not predict that countries that have reformed will on average have faster growth than countries that have not. What the model does predict—and the Baker Hypothesis implicitly claims—is the following: If a given country implements and maintains certain economic reforms, then its gross domestic product (GDP) will grow faster after the reform than it did prior to implementation. The period of faster growth will persist until the country has completed its transition to the new, higher level of total factor productivity induced by the reform. Once the transition is over, the country, now at a permanently higher level of GDP, will revert to its pre-reform, steady state rate of growth. Although such a transition can take decades to complete, the calculations in Henry (2007, pp. 898–899) demonstrate (in the context of capital account liberalization) that the average deviation of GDP growth from its steady-state value in the first five years after the policy change is 2.5 times larger than the average deviation in years six through 30. In addition, given the magnitude of standard errors associated with cross-sectional regression estimates, the average growth deviation during the first five years of transition will be statistically as well

as economically significant, while the average deviation in years six through 30 will not. In an example in the same spirit, Easterly (1996) illustrates that cross-sectional regressions of inflation on growth have little power to discern the true impact of stabilizing inflation. Finally, Wacziarg and Welch (2008) show that the positive relationship between growth and open trade regimes documented by Sachs and Warner (1995) is economically larger and statistically more robust when, as suggested by Henry (2007), one uses a panel data, event-study approach to test explicitly for the presence of a temporary growth acceleration, both on impact and in the immediate aftermath of major policy changes.

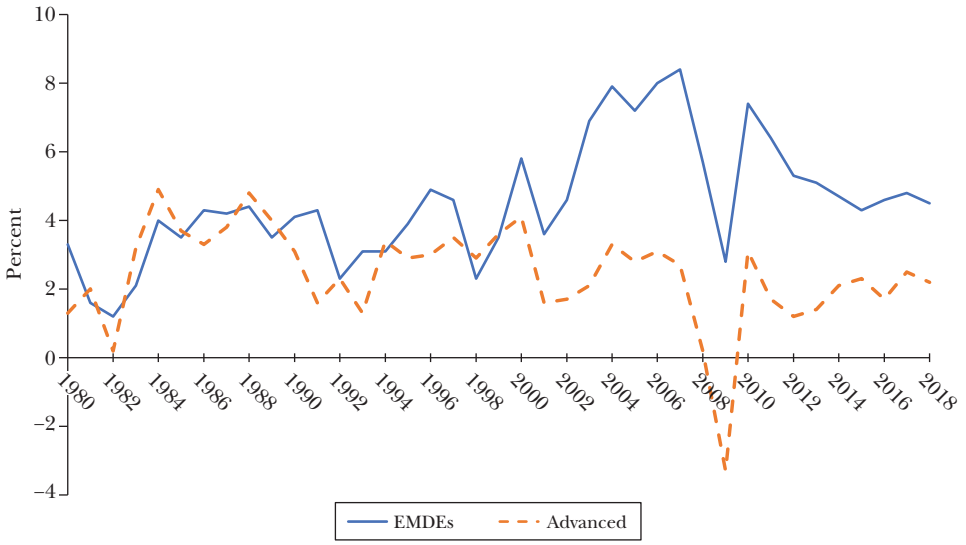
This article takes seriously the time series, country-specific predictions of the Solow model. It does so by focusing on the widespread, if uneven, adoption of a set of policy reforms—commonly referred to as the "Washington Consensus mantra of stabilize, liberalize, and privatize" (Gertz and Kharas 2019)—by emerging and developing economies in the late 1980s and into the early-to-mid 1990s. In particular, the key policies are 1) stabilization of inflation; 2) freer trade; 3) increased openness to flows of foreign investment; and 4) an expanded role of the market in producing and allocating goods and services through privatization. Looking at these changes, in turn, provides a set of policy experiments that enable us to examine whether the time paths of GDP growth associated with these reforms refute or support Baker's implicit "if-then" claim.

Setting the Stage

Figure 1 uses the IMF's weighted average of real GDP growth for all emerging and developing economies, as well as that of all advanced countries, to set the stage for the country-specific discussion to follow.[1] Through the 1980s, the average growth rate of real GDP for the emerging and developing economies was quite similar to that of the advanced economies even though theory predicts, all else equal, that the emerging and developing economies should have been experiencing catch-up growth and therefore expanding more rapidly than the advanced economies. Because of their problems with debt and inflation, however, all else was not equal in the emerging and developing economies until many of their leaders initiated economic reforms. The reform process that had been set in motion by Baker's 1985 speech was pushed forward by the fall of the Berlin Wall in 1989 and was reinforced by the implementation of debt relief agreements under the Brady Plan in the early 1990s (Williamson 2004). After peaking in 1993, the dramatic and permanent fall in the IMF's weighted average of inflation in the emerging and developing economies, shown in Figure 2, provides a salient indicator of the meaningful, if imperfect, shift that took hold in the economic policies and priorities of much of the developing world following the Brady Plan debt relief agreements.

[1] One country in our sample, South Korea, would have been classified as an emerging and developing economy in 1980 but is now "advanced."

Figure 1
Emerging and Developing Economies Grew Faster after They Implemented Reforms



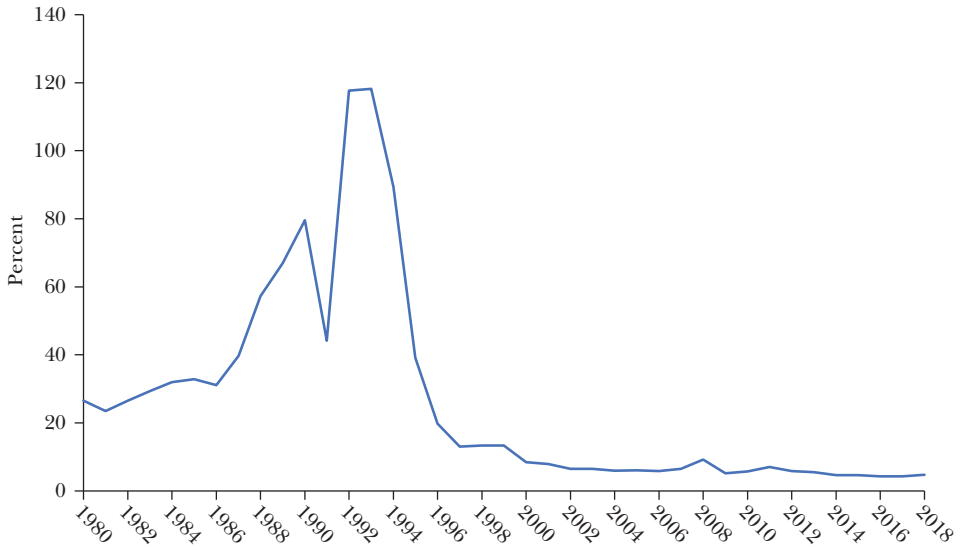
Note: Figure 1 presents the IMF's weighted average of real GDP growth for all emerging and developing economies, as well as that of all advanced economies for comparison. EMDE stands for "emerging market and developing economies." Details of data and calculations are in the online Appendix available with this article at the *JEP* website.

As countries stabilized inflation and pursued structural reforms, catch-up growth ensued. Going back to Figure 1, and taking 1993 as a proxy date for the turnaround, we calculate that from 1994 to 2018, real GDP in the emerging and developing economies grew by an average of 5.2 percent per year versus the 3.3 percent rate at which it expanded from 1980 to 1993.[2] For a country whose population increases at the rate of 1 percent per year, 3.3 percent GDP growth means that its per capita income doubles once every 30 years; with 5.2 percent GDP growth, the same country's standard of living doubles in just 16 years. The growth acceleration in emerging and developing economies was not driven by China. Indeed, China's average rate of growth actually slowed over the period in question—from 9.9 percent between 1980 and 1993 to 9.6 percent from 1994 to 2018. Thus, the 1.9 percentage-point increase in the average growth rate of emerging and developing economies is not a statistical artifact of China's economic performance.

Also, contrary to the popular zero-sum media narratives that faster growth in poor countries harms rich countries (summarized in Krugman 1994), there is no evidence that rising standards of living in the emerging and developing economies came at the expense of "first world" prosperity. The advanced countries grew by

[2] Shifting the proxy date for the turnaround by a year or two in either direction has a de minimus impact on our calculation.

Figure 2
The Timing of the Permanent Fall in Inflation in Emerging and Developing Economies



Note: Figure 2 shows that after peaking in 1993, there was a dramatic and permanent fall in the IMF's weighted average of inflation in emerging and developing economies. Details of data and calculations in the online Appendix.

2.9 percent per year from 1980 to 1993; from 1994 onward, they continued growing at approximately the same rate (excluding the period of the global financial crisis that originated in the developed world).

The largely unchanged long-term growth performance in rich countries additionally suggests that the accelerated rise of living standards in poor ones was not driven by an aggregate shock to the global economy, but rather by factors specific to the emerging and developing economies. Population expansion provides one potential alternative explanation to reforms, as vast supplies of low-cost labor in the rural and informal economies of poor countries surely played a role in sustaining the growth process (as in a Lewis-style 1954 growth model). But from 1994 to 2018, there was no change in the demographics of the developing world to suggest that an increase in the growth rate of its working age population was responsible for the growth acceleration. In fact, from 1994 to 2018, the growth rate of the working age population in Asia and Latin America was actually decreasing (and was roughly constant in Africa), even as the growth rate of real GDP for emerging and developing economies was rising.[3]

[3]Non-demographic factors may also explain the growth acceleration in emerging and developing economies. While this essay is not the place for an econometric examination of the extent to which the Baker

Instead, the proximate cause of the growth acceleration in emerging and developing economies was, indeed, the array of country-specific economic reforms pushed forward by the Brady Plan. In countries that implemented and maintained reforms, the level of productivity rose. With wages remaining flat for an extended period of time due to a highly elastic supply of labor, owners of capital had a persistent incentive to invest, triggering, in turn, a cycle of sustained profitability and expanding demand for previously underemployed workers.

While the data on economic outcomes speak clearly in retrospect, the path to meaningful reforms that brought them about was slow, rocky, and non-linear. In Baker's (1985) speech, he failed to say that his remarks provided a compass, not a map. Postulating that developing countries would grow faster if they stabilized and traded more with the rest of the world was one thing. Charting a course from the universe of potential policy changes he described to higher standards of living was quite another. The second step required, for each nation, a sustained commitment to a pragmatic growth strategy, consisting of an optimal mix of country-specific, efficiency-enhancing policy changes (Henry 2013). Indeed, one might say that the empirical success of the Baker Hypothesis, which conjectured what reforms would make economies grow, stands in sharp contrast to the failure of the Baker Plan, which did not articulate a realistic strategy for how leaders could actually bring about the subset of reforms best suited to their countries. Starting from the creation of macroeconomic stability, a condition without which there is no sustained growth (Commission on Growth and Development 2008), the rest of this article provides a country-specific, time-series assessment of the economic reform process.

Stabilization of Inflation

The intellectual justification for Baker's (1985) call to reduce inflation flows from the reality that stabilizing high inflation raises productivity because, among other reasons, stabilization reduces the variance of the aggregate price level as well as the variance of relative prices. The variability of the aggregate price level matters, because greater variability of inflation increases the likelihood of bouts of high and unexpected inflation (Ha, Kose, and Ohnsorge 2019; IMF 2001). High inflation is not neutral and therefore creates relative price distortions that reduce the quality of the signal that individual prices provide to producers about the profitability of goods and services, thereby increasing uncertainty about profitability and reducing the incentive to produce and invest (Andrés and Hernando 1999). Because unexpected inflation helps borrowers and hurts lenders, fear of unexpected inflation

Hypothesis stands up to a range of alternative explanations, the central conclusions about reforms and growth suggested by the figures we show withstand empirical scrutiny elsewhere. For evidence on the impact of stabilization see Easterly (1996) and Henry (2002); on trade liberalization, Wacziarg and Welch (2008), Estevadeordal and Taylor (2013), and Irwin (2019); on capital account liberalization, Henry (2000a), Chari and Henry (2004), Henry (2007), and the references therein.

in non-inflation-indexed environments may discourage lenders from entering into long-term contracts, again with negative attendant consequences for production and investment.

In his 1985 speech, Baker did not specify the level at which he and his US Treasury colleagues considered inflation to be "high," but in keeping with previous work we define high inflation as annual consumer price index inflation that is 40 percent or more; "moderate" inflation is less than 40 percent but greater than or equal to 10 percent; and "low" inflation is less than 10 percent (Dornbusch and Fischer 1993; Fischer 1993; Easterly 1996; Bruno and Easterly 1998).

Next, we use these definitions to determine the year in which a given country stabilized inflation in the following manner. First, we gather the country's annual rates of consumer price inflation from World Bank data and construct a time series of its three-year moving average of inflation. Second, starting from the initial year of the series, we classify the country's level of inflation in accordance with the first instance in which the country experiences high or moderate inflation for five or more consecutive years. Third, we identify when the country's classification shifts into the next lowest group (for example, from "moderate" to "low," or from "high" to "moderate") for five or more consecutive years (again, using a three-year moving average for each year). We define the country's "stabilization year" as the peak-inflation year identified by our procedure. Finally, we classify each country's stabilization episode as "high" if the stabilization began from an inflation peak that was "high," and "moderate" if its stabilization episode began from a peak that was "moderate." Our procedure yields 25 "high" and 28 "moderate" inflation-stabilization episodes in emerging and developing economies, for a total of 53 stabilization episodes. The number of episodes is less than the number of emerging and developing economies listed by the IMF because some countries did not have a stabilization. Also, because we seek to examine the growth rate of real GDP in the decade before and after stabilization, we dropped 38 countries for lack of data.[4]

Table 1 summarizes the 53 inflation stabilization episodes. Panel A indicates that when it comes to stabilizing high inflation, the average year of stabilization across all regions is 1992. Among regions, Latin America has the greatest frequency of high inflation stabilizations, with 11 of the 25 episodes and an average peak inflation rate of almost 1,000 percent. There were eight high inflation episodes in Africa. For stabilizations of moderate inflation, as shown in Panel B, Africa contains 14 of the 27 episodes, the average stabilization year is 1990 (the median is 1989), and the average level of moderate inflation at the peak was 22 percent. In Latin America, the seven cases of moderate inflation peaked at an average rate of 27 percent in 1996. Turning to Asia, it is notable that South Korea stabilized moderate inflation in 1982, much earlier than the vast majority of the other inflation stabilization episodes in emerging and developing economies.

[4]The online Appendix available with this paper at the *JEP* website lists the 38 countries we dropped as well as all of those classified as emerging and developing economies by the IMF.

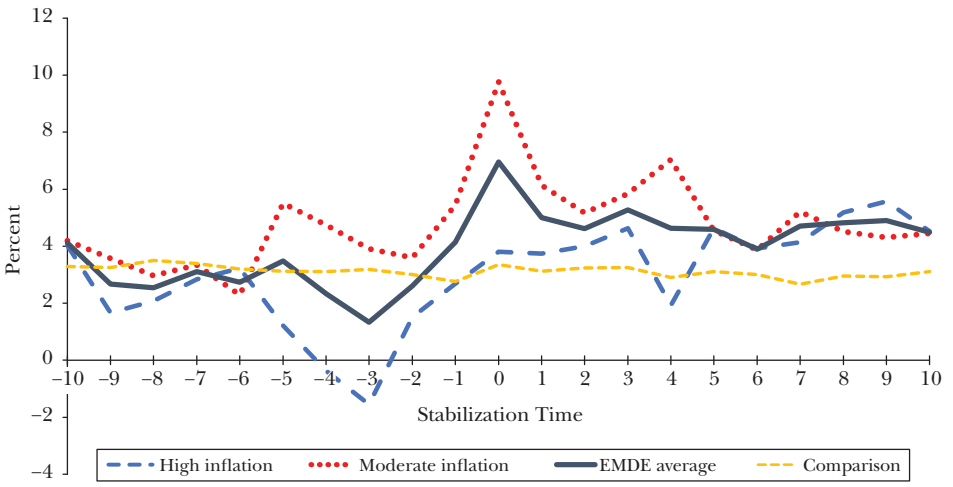
Table 1
Inflation Stabilization Episodes by Geography

Africa		*Asia*		*Latin America*		*Eastern Europe*	
Country (Year)	*Inflation (%)*	*Country (Year)*	*Inflation (%)*	*Country (Year)*	*Inflation (%)*	*Country (Year)*	*Inflation (%)*
Panel A. High inflation							
Angola (1997)	2,587	Lao PDR (2000)	81	Bolivia (1987)	4,435	Albania (1995)	111
Congo, Dem. Rep. (1995)	9,963	Mongolia (1996)	118	Brazil (1995)	1,651	Bulgaria (1998)	414
Ghana (1984)	87	Syrian Arab Rep. (1989)	43	Chile (1976)	410		
Guinea-Bissau (1994)	58	Turkey (1997)	85	Costa Rica (1984)	53		
Malawi (1997)	51			Dominican Republic (1992)	46		
Nigeria (1996)	62			Ecuador (2002)	62		
Sudan (1994)	114			Jamaica (1994)	50		
Zambia (1994)	148			Mexico (1989)	110		
				Peru (1991)	3,849		
				Suriname (2002)	65		
				Uruguay (1992)	98		
Mean (1994)	1,634	Mean (1996)	82	Mean (1991)	984	Mean (1997)	262
Median (1995)	100	Median (1997)	83	Median (1992)	98	Median (1997)	262
Number of countries	8	Number of countries	4	Number of countries	11	Number of countries	2
Panel B. Moderate inflation							
Botswana (1994)	14	Kyrgyz Rep. (2000)	23	Colombia (1993)	28	Papua New Guinea (2001)	14
Cote d'Ivoire (1980)	19	Myanmar (1976)	27	Dominican Republic (2005)	28		
Algeria (1995)	27	Myanmar (1999)	32	Guatemala (1992)	28		
Egypt, Arab Rep. (1989)	20	Pakistan (1976)	23	Haiti (2006)	21		
Gambia, The (1988)	32	Philippines (1992)	14	Honduras (1997)	25		
Equatorial Guinea (1997)	19	Korea (1982)	22	Paraguay (1992)	29		
Kenya (1995)	34			El Salvador (1988)	26		
Kenya (2008)	16						
Rwanda (1976)	23						
Madagascar (1997)	35						
Senegal (1985)	13						
Eswatini (1988)	15						
Seychelles (1975)	21						
South Africa (1988)	17						
Mean (1990)	22	Mean (1988)	24	Mean (1996)	26	Mean (2001)	14
Median (1989)	19	Median (1987)	23	Median (1993)	28	Median (2001)	14
Number of countries	14	Number of countries	6	Number of countries	7	Number of countries	1

Note: Table 1 summarizes the 53 inflation stabilization episodes. Panels A and B list "high" and "moderate" inflation countries by world region, the stabilization year, and the peak inflation rate. High inflation is an annual consumer price index inflation that is 40 percent or more; "moderate" inflation is less than 40 percent but greater than or equal to 10 percent; and "low" inflation is less than 10 percent. Using a time series of a three-year moving average of inflation, we classify a country as having high, moderate, or low inflation at the start of the three-year moving average. Starting from the initial year of the series, we identify the first instance in which the country experiences a level of inflation that shifts its classification into the next lowest group (for example, from "moderate" to "low," or from "high" to "moderate") for five or more consecutive years. We define the country's "stabilization year" as the peak-inflation year identified by our procedure. We classify each country's stabilization episode as "high" if the stabilization began from an inflation peak that was "high," and "moderate" if its stabilization episode began from a peak that was "moderate." Our procedure yields 25 "high" and 28 "moderate" inflation-stabilization episodes in emerging and developing economies.

Figure 3

Emerging and Developing Economies Grew Faster after They Stabilized Inflation



Note: Figure 3 plots, in event time with the year of inflation stabilization as zero, the average growth rate of real GDP surrounding the 53 country-inflation-stabilization episodes. The figure also includes a plot in stabilization time of the average growth rate of real GDP for a comparison group of countries. For a given emerging market and developing economy (EMDE) inflation stabilization episode (for example, Brazil 1995), the comparison group consists of all countries that meet the World Bank's income threshold for being classified as "advanced." We then take as the comparison-group growth series for the given episode, the World Bank's (weighted) average growth rate of advanced economies for each of the years in the interval [–10, 10] (e.g., [1983, 2003] for Brazil). Proceeding in identical fashion for each emerging-and-developing-economy episode, we construct 53 series of comparison-group growth rates. The "comparison" line in Figure 3 is the (unweighted) average of these 53 series. The figure also plots the average growth rate of GDP for the "high" and "moderate" inflation stabilization episodes. Details of data and calculations are in the online Appendix.

Figure 3 uses IMF data to plot, in event time with the year of inflation stabilization as zero, the average growth rate of real GDP surrounding the 53 country-inflation-stabilization episodes. The figure also includes a plot in stabilization time of the average growth rate of real GDP for a comparison group of countries. We construct the comparison group as follows. For a given emerging and developing economy inflation stabilization episode (say, Brazil 1995), the comparison group consists of all countries that meet the World Bank's income threshold for being classified as "advanced."[5] We then take as the comparison-group growth series for the given episode, the World Bank's (weighted) average growth rate of advanced economies for each of the years in the interval [–10, 10] (for

[5]We considered including in the comparison group only those advanced economies that had "low" inflation for at least ten years prior to the year of the emerging and developing economy stabilization episode, but the number of advanced economies across all 53 episodes that did not meet the low-inflation threshold was negligible.

Brazil 1995, this would be the interval [1985, 2005]). Proceeding in identical fashion for each emerging-and-developing-economy episode, we construct 53 series of comparison-group growth rates. The "comparison" line in Figure 3 is the (unweighted) average of these 53 series.

The average annual growth rate of real GDP ten years after the onset of stabilization is 4.9 percent versus 2.9 percent in the ten years prior, 37 of 53 countries have a post-stabilization growth rate of GDP that is higher than their country-specific, pre-stabilization growth rate, and the growth collapse in the years immediately preceding stabilization mirrors the findings of Bruno and Easterly (1998).[6] Consistent with previous work on the costs and benefits of stabilization (Easterly 1996; Henry 2002), the growth increase in the aftermath of stabilizing high inflation is larger than in the case of stabilizing moderate inflation. For the "high" episodes, the average annual growth rate of GDP rises from 1.6 percent prior to stabilization to 4.2 percent after, an increase of 2.6 percentage points per year. Concurrent events temper interpretation of the magnitude of the impact, but the directional effect is robust. Twenty-one of 25 countries have an average post-stabilization rate of growth that exceeds their country-specific, pre-stabilization average, and 21 of 25 countries have a median post-stabilization rate of growth that exceeds their country-specific, pre-stabilization median. The four countries that do not experience an increase in average or median growth after stabilization are Brazil, Guinea Bissau, Jamaica, and Malawi.

The average annual growth rate of GDP also rises for the "moderate" episodes—from 4.05 percent to 5.52 percent—but the increase of 1.47 percentage points is a little less than three-fifths the size of that in the "high" episodes, and the pattern of increase is less consistent. Sixteen of 28 countries have an average post-stabilization rate of growth that exceeds their country-specific, pre-stabilization average, and 16 of 28 countries have a median post-stabilization rate of growth that exceeds their country-specific, pre-stabilization median. The trajectory of the comparison group is flat.

Liberalization of Trade

Because stable and predictable inflation increases the informativeness of prices and improves the efficiency of resource allocation, there is broad agreement that stabilizing inflation—and therefore the macroeconomic environment more generally—is a necessary condition for a country to maximize the benefits of opening up its economy to trade and capital flows from the rest of the world (Fischer 1986, 1987; Mathieson and McKinnon 1981; McKinnon 1984; Michalopolous 1987; Sachs 1988). At the time of Baker's (1985) speech, however, there was considerably less agreement about whether the benefits of a country opening up would outweigh the

[6] For a list of the 37 countries and other details of these calculations, see the online Appendix.

costs (Sachs 1987). Baker (1985) and his Treasury colleagues had no such qualms and Baker (p. 209) argued: "For those countries which have implemented measures to address the imbalances in their economies, a more comprehensive set of policies can now be put in place . . . We believe that such institutional and structural policies should include . . . market-opening measures to encourage foreign direct investment and capital inflows, as well as to liberalize trade."

Wacziarg and Welch (2008, building on Sachs and Warner 1995) carefully construct a comprehensive collection of country-specific trade liberalization dates. From the Wacziarg and Welch list of 98 advanced and developing countries that have liberalized trade, we culled the dates of the 72 countries in their sample that were classified as developing countries at the time of Baker's (1985) speech. Of these 72 countries, 64 had a sufficiently long time series on real GDP growth to be included in our analysis.[7] Table 2 summarizes these 64 episodes. From a temporal perspective, most countries liberalized trade in the early 1990s, and the average trade liberalization year for the entire sample is 1990 (median of 1991). From a geographic standpoint, Africa had the largest number of countries that liberalized trade, with 26.

Korea in 1968 stands out as an early liberalizer of trade, just as it did as an early stabilizer of inflation. Korea's early mover status on a subset of structural economic reforms is somewhat at odds with the narrative that places the roots of Korea's successful growth experience in government interventionism. As Amsden (1989, p. 80) writes: "Every major shift in industrial diversification in the 1960s and 1970s was instigated by the state." But in our view, the key input into Korea's economic transformation was less an ideological tilt toward *dirigisme* than it was a commitment by the state to a pragmatic growth strategy that empowered Korean enterprises to become more active and effective participants in the world market. The Korean approach to trade liberalization, along with that of Singapore and Taiwan, contained two critical elements that constitute, as it were, a test of the Baker Hypothesis before Baker.

First, Korea, Singapore, and Taiwan all stabilized inflation before pursuing trade liberalization and remained vigilant about maintaining macroeconomic stability. Korea, although we do not have the data on consumer price inflation to detect it in our stabilization algorithm, experienced hyperinflation in the 1950s—during and after the Korean War—that it reduced to moderate inflation by 1960. Like Korea, Taiwan also experienced hyperinflation—during the Chinese Civil War—but stabilized inflation by 1951 (Sachs 1987). As for Singapore, with the exception of a temporary spike in 1973 and 1974 due to the oil-price shock, the country has had low inflation since its independence in 1965.

Second, by the end of the 1960s, Korea, Singapore, and Taiwan had all rejected an import substitution strategy for promoting growth and instead embraced a sustained commitment to growth strategies that relied on both imports and exports

[7] The list of 72 countries and dates, along with more details of the analysis, is available in the online Appendix.

Table 2
The Frequency of Trade Liberalization Episodes Varies by Geography and Was Concentrated in the 1990s

Africa *Country (Year)*	*Asia* *Country (Year)*	*Latin America* *Country (Year)*	*Eastern Europe* *Country (Year)*
Benin (1990)	Bangladesh (1996)	Argentina (1991)	Albania (1992)
Botswana (1979)	Jordan (1965)	Bolivia (1985)	Bulgaria (1991)
Burkina Faso (1998)	Korea (1968)	Brazil (1990)	Georgia (1996)
Burundi (1999)	Nepal (1991)	Chile (1976)	Hungary (1990)
Cabo Verde (1991)	Pakistan (2001)	Colombia (1986)	Montenegro (2001)
Cameroon (1993)	Philippines (1988)	Costa Rica (1986)	Poland (1990)
Cote d'Ivoire (1994)	Sri Lanka (1991)	Dominican Republic (1992)	Romania (1992)
Egypt (1995)	Tajikistan (1996)	Ecuador (1991)	Serbia (2001)
Ethiopia (1996)	Turkey (1989)	El Salvador (1989)	
Gambia (1985)		Guatemala (1988)	
Ghana (1986)		Guyana (1988)	
Guinea (1986)		Honduras (1991)	
Guinea-Bissau (1987)		Jamaica (1989)	
Kenya (1993)		Mexico (1986)	
Madagascar (1996)		Nicaragua (1991)	
Mali (1988)		Panama (1996)	
Mauritania (1995)		Paraguay (1989)	
Mauritius (1968)		Peru (1991)	
Morocco (1984)		Trinidad and Tobago (1992)	
Mozambique (1995)		Uruguay (1990)	
Niger (1994)		Venezuela (1996)	
Sierra Leone (2001)			
South Africa (1991)			
Tanzania (1995)			
Tunisia (1989)			
Uganda (1988)			
Mean (1990)	Mean (1987)	Mean (1989)	Mean (1994)
Median (1991)	Median (1991)	Median (1990)	Median (1992)
Number of countries 26	Number of countries 9	Number of countries 21	Number of countries 8

Note: Table 2 summarizes 64 trade liberalization episodes. We use information from Wacziarg and Welch (2008, building on Sachs and Warner 1995) who construct a comprehensive collection of country-specific trade liberalization dates. From the Wacziarg and Welch list of 98 advanced and developing countries that have liberalized trade, we culled the dates of the 72 countries in their sample that were classified as developing countries at the time of Baker's (1985) speech. Of these 72 countries, 64 had a sufficiently long time series on real GDP growth to be included in our analysis.

(Commission 2008). While Korea did not fling its economy wide open—the country retained high import tariffs on a wide range of items from agricultural products

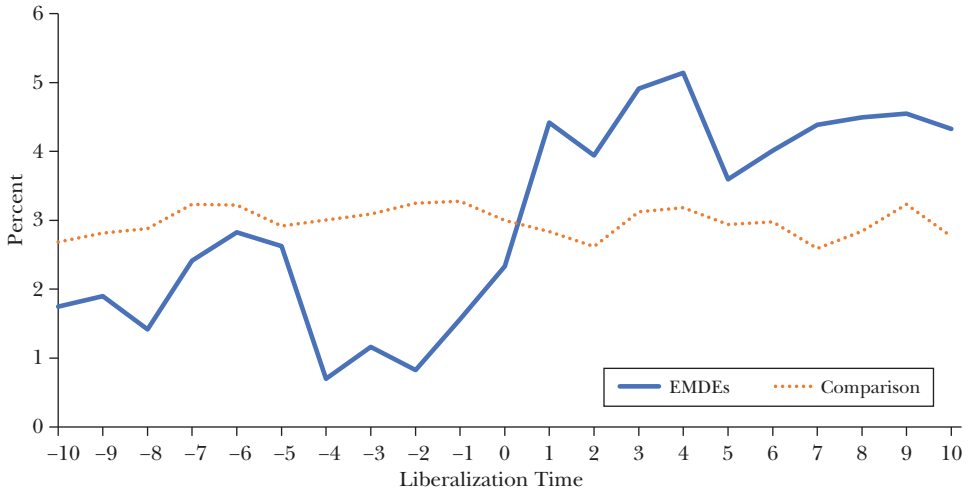
to computer equipment—the authorities acknowledged the necessity of certain imported foreign goods and acted accordingly.

As an important complement to the import liberalization agenda, in 1964 Korean leaders reduced the fiscal deficit and devalued the won by almost 100 percent in order to increase export profitability (Dornbusch and Park 1987). But Korea's approach was not mercantilist. Although exports rose from 4.8 percent of GDP in 1963 to 34 percent of GDP in 1980, imports increased from 15.9 percent of GDP to 41.4 percent of GDP over the same time period (Krueger 1995, Table 1.4). In other words, while the maxi-devaluation achieved the goal of increasing exports relative to imports—exports rose by a factor of 7.1, imports by a factor of 2.6—it did not undermine the integral role of foreign goods in the country's development. In fact, from 1965 to 1990, Korea's real GDP per capita grew by 7.1 percent per year, with the country running trade deficits for almost the entire period (Krueger 1995, Tables 1.1 and 1.3). In the case of Taiwan, import tariffs were similarly reduced, and a large number of items—intermediate capital inputs, in particular—were removed from the import control list. As in Korea, Taiwanese officials also corrected the overvaluation of its currency. They did this by: 1) devaluing the New Taiwan Dollar by between 50 and 80 percent from 1958–1961, depending on the type of transaction; and 2) unifying the exchange rate in 1963. Taiwanese officials also established export processing zones and passed a law in 1960 to permit direct investment by foreign and overseas Chinese capital (Jao 1976). Like Taiwan, Singapore also chose to encourage foreign direct investment as it switched to export-led growth (Menon 2015).

Turning from East Asia back to the broader developing world, Figure 4 uses IMF data to plot, in trade liberalization time, the (unweighted) average growth rate of GDP for the 64 emerging and developing economies that undertook trade liberalization. The figure also includes a plot in liberalization time of the growth rate of GDP for a comparison group of countries. We construct the comparison group as follows. For a given emerging-and-developing-economy trade liberalization episode (say, Egypt 1995), the comparison group consists of all advanced economies in the IMF data that, per Wacziarg and Welch (2008), were classified as having "free trade" at least ten years prior to the year of the emerging-and-developing-economy trade liberalization episode. We then use the comparison group of advanced economies to construct the comparison growth series for the given episode as the (unweighted) average rate of growth for each of the years in the interval [−10, 10] across all advanced economies in the group (say, the interval [1985, 2005] for Egypt). Proceeding in identical fashion for each emerging-and-developing-economy trade liberalization episode, we construct 64 series of comparison-group growth rates. The line "comparison" in Figure 4 is the unweighted average of these 64 series.

For the 10-year period before trade liberalization, the average growth rate of real GDP in the 64 emerging and developing economies was 1.72 percent. The average growth rate of real GDP in these economies for the 10-year post-liberalization period was 4.38 percent. The 2.66 percentage-point increase in the average growth rate of GDP in the emerging and developing economies, again tempered by concurrent events, is not driven by outliers but rather a consistent pattern of

Figure 4
Emerging and Developing Economies Grew Faster after They Liberalized Trade



Note: Figure 4 plots in trade liberalization time, the (unweighted) average growth rate of GDP for the 64 emerging market and developing economies (EMDEs) that undertook trade liberalization. The figure also includes a plot in liberalization time of the growth rate of GDP for a comparison group of countries. For a given emerging- and-developing-economy liberalization episode (for example, Egypt 1995), the comparison group consists of all advanced countries in the IMF data that, per Wacziarg and Welch (2008), were classified as having "free trade" at least ten years prior to the year of the emerging-and-developing-economy liberalization episode. We then use the comparison group of advanced countries to construct the comparison growth series for the given episode as the (unweighted) average rate of growth for each of the years in the interval [–10, 10] across all countries in the group (for example, the interval [1985, 2005] for Egypt). Proceeding in identical fashion for each emerging-and-developing-economy trade liberalization episode, we construct sixty-four series of comparison-group growth rates. The line "comparison" is the unweighted average of these 64 series. Details of data and calculations in the online Appendix.

higher growth after opening. Of the 64 countries in the sample, 52 have a post-trade liberalization growth rate that exceeds their country-specific, pre-liberalization average. The median post-liberalization growth rate exceeds the country-specific, pre-liberalization median in 53 cases. The trajectory of the comparison group is flat.

Liberalization of the Capital Account

Baker's (1985) case for developing countries opening to foreign investment rests on standard neoclassical theory, in which liberalizing the capital account facilitates a more efficient international allocation of resources. Specifically, savings flow from capital-abundant developed countries, where the return on capital is low, to capital-scarce developing countries where the return on capital is high. The flow of savings into the developing countries reduces their cost of capital, triggering a

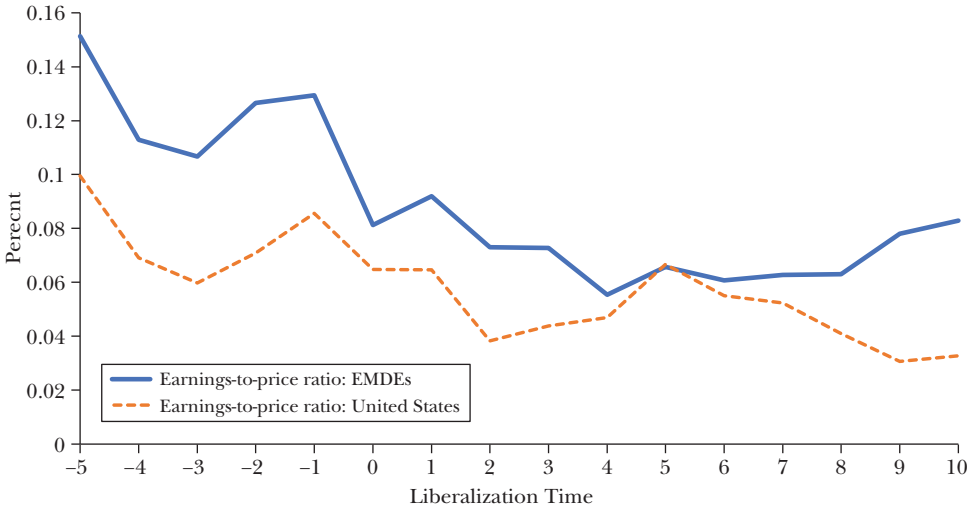
temporary increase in investment and growth that permanently raises their standard of living (Fischer 2003; Krueger 1988; Obstfeld 1998).

The national stock market's earnings-to-price ratio, the aggregate earnings yield, is the average cost of equity capital for all publicly traded firms in a country. The aggregate earnings yield, therefore, provides the broadest visible proxy for the rate of return that owners of capital require to reinvest their profits in the local economy instead of allocating them elsewhere or increasing consumption. In turn, the aggregate earnings yield equals the risk-free interest rate plus the equity-risk premium. In theory, prior to liberalization of the capital account, the risk-free rate for a given country is determined domestically by the local supply of savings and demand for investment; the country's pre-liberalization, equity-risk premium is the domestic price of risk (required return per unit of variance) multiplied by the quantity of risk (the variance of aggregate market returns). After liberalization, the country's capital market is integrated with the world capital market; therefore, post liberalization, the risk-free rate is the world interest rate, and the equity premium is the world price of covariance risk multiplied by the covariance of local market returns with global market returns. Because the world risk-free rate is typically lower than the risk-free rate for emerging and developing economies, and the variance of emerging stock returns is greater than their covariance with world stock returns (Chari and Henry 2004; Stulz 1999), it is reasonable to expect liberalization to reduce the aggregate earnings yield.

We define "capital account liberalization" as the first point in time that a government permits foreigners to purchase shares of publicly listed corporations. This may seem to be a limited form of opening an economy to international capital flows, but the easing of foreign ownership restrictions on domestic stocks, in addition to enabling flows of portfolio equity, played a significant role in facilitating foreign direct investment and privatization (Edwards 1995, chapter 6). Using dates from Chari, Henry, and Sasson (2012), Figure 5 plots, in liberalization time, the average value of the earnings yield of the 18 emerging and developing economies for which there is information on both liberalization dates and the earnings-to-price ratio as a basis for calculating the aggregate earnings yield. Again, the online Appendix contains a list of the 19 countries for which we have both liberalization dates and earnings yields, as well as the additional six countries for which we have dates—again from Chari, Henry, and Sasson (2012)—but no yields. The numbers are annual, and the plot starts at year −5 because of data limitations (the average and median liberalization date is 1989, and there are only three countries with data on earnings yields in year −10, none of which are in Latin America). As a point of comparison, and a comparison group, Figure 5 also plots the US earnings yield to which we assign a year "0" of 1989 to match the average liberalization date of the emerging and developing economies. Two aspects of the figure are noteworthy.

First, during the process of capital account liberalization and its aftermath, the average earnings yield of emerging and developing economies falls sharply and then gradually converges to that of the US economy. As shown in Figure 5, on impact—that is, between year −1 and year 0—the average earnings yield in emerging and

Figure 5
The Cost of Equity Capital in Emerging and Developing Economies Fell When They Eased Restrictions on Foreign Ownership of Domestic Stocks



Note: Figure 5 plots, in liberalization time, the average value of the earnings yield of the 18 emerging market and developing economies (EMDEs) for which there is information on both liberalization dates and the earnings-to-price ratio as a basis for calculating the aggregate earnings yield. The numbers are annual, and the plot starts at year -5 because of data limitations. As a point of comparison, Figure 5 also plots the US earnings yield to which we assign a year "0" of 1989 to match the average liberalization date of the emerging and developing economies. Details of data and calculations in the online Appendix.

developing economies falls from 12.9 percent to 8.1 percent, a drop of 480 basis points in a single year. This decline is not the result of a few outliers, but instead a consistent fall in the cost of equity capital. Only four of the 19 countries—Chile (1987), India (1986), Malaysia (1987), and Thailand (1987)—do not on impact experience a fall in their earnings yield. The US earnings yield also falls during the liberalization window but by a smaller amount, 210 basis points from 8.6 percent to 6.5 percent. The gap between the earnings yield of the emerging and developing economies and that of the United States continues to narrow in the aftermath of liberalization, converging to zero in year 5.

Second, except for the rise in year 9 associated with the timing of the 1997–98 Asian financial crisis, the fall in earnings yields for emerging and developing economies appears to be permanent. The average yield for the 19 countries in the five years prior to liberalization of 12.5 percent drops to an average of 7.1 percent in the ten years after liberalization—a decrease of 540 basis points. In the case of the US equity market, the average earnings yield is also lower in the post-liberalization period than it was in the pre-liberalization period—4.7 percent versus 7.9 percent—but the decline in the average yield in emerging and developing economies is 220 basis points larger. The pattern of a longer-run post liberalization cost of equity

capital that is lower than the pre-liberalization cost is also extremely consistent. Of the 19 countries, South Africa with a 30-basis-point increase is the only country whose average cost of capital in the 10-year post-liberalization period is not lower than its five-year, country-specific, pre-liberalization average. South Africa is also the only country whose median post-liberalization cost of capital is not lower than its country-specific, pre-liberalization median.

The fall in the required rate of return for stocks, in conjunction with the onset of inflation stabilization and trade liberalization, provides a plausible, if admittedly oversimplified, explanation of the growth acceleration that took hold in the developing world in the early and mid-1990s. By reducing inflation to provide stability and reduce uncertainty, as well as opening the economy to increase the supply of savings and allow greater diversification of risk, the combination of macroeconomic stabilization and capital account liberalization reduced the cost of equity capital in emerging and developing economies. By tilting domestic output in the direction of comparative advantage and raising productivity, trade liberalization raised the aggregate rate of return on investing in capital. Falling costs of capital in conjunction with higher prospective returns to property, plants, and equipment provided a strong incentive to increase investment, and many countries in the developing world did, in fact, experience higher growth rates of capital, wages, and GDP following major reforms (Chari, Henry, and Sasson 2012; Chari and Henry 2008; Henry 2000b; Henry 2007).

In addition to giving emerging and developing economies access to a larger pool of savings, opening their stock markets to foreigners enabled developing nations to reduce their reliance on debt, which requires payments that are invariant to the borrower's circumstances, because they could instead resort to foreign direct investment and portfolio equity as alternative sources of capital. Baker's (1985, 210) speech mentioned the benefits of foreign equity financing as a complement to debt, but his remarks did not address a critical source of debt bias in the international financial system: implicit subsidies to *suppliers* of debt capital. Lenders to emerging and developing economies from the "G-7 countries" (Canada, France, Germany, Italy, Japan, United Kingdom, and United States) resort to G-7 courts in the event of debt disputes, but there is no such recourse for G-7 holders of emerging market equity (Bulow 2002; Rogoff 1999). Failure to address the debt bias left emerging and developing economies vulnerable in the future to the excessive reliance on leverage that lay at the heart of the 1980s debt crisis. Little surprise, then, that an overload of short-term, dollar-denominated debt was the proximate cause of both the 1994 Mexican crisis and the 1997–1998 Asian crisis (Feldstein 2002). Said another way, it is not capital account liberalizations per se that cause crises. The danger, instead, lies with liberalizations that ease restrictions on foreign borrowing (bonds and bank loans) without first implementing prudential regulations that guard against the pitfalls of leverage. Policy changes that grant legal protections for equity financing of investment in developing countries that are as strong as the protections in place for debt financing would mitigate debt bias and reduce the frequency of future financial crises in emerging markets.

There is an irony here. The aversion of developing country leaders in the 1970s to allowing foreigners to purchase shares in their countries' corporations created an excessive reliance on leverage, that when combined with adverse shocks led to a debt crisis, which, in turn, left them little choice but to open their equity markets to facilitate foreign direct investment—including the wave of privatizations that began in Latin America and spread to the former Soviet Union and Eastern Europe.

Privatization of State-Owned Enterprises

With the sale of British Telecom by the Thatcher government in 1984, the term "privatization" entered the everyday lexicon of modern economics, but in the aftermath of Baker's (1985) speech, the trend of selling state-owned enterprises took hold in the developing world. The output of state-owned enterprises as a share of GDP fell from a peak of 10 percent in 1986 to 4 percent in 1995 for upper-middle income developing countries, 12 percent in 1982 to less than 6 percent in 1995 for lower-middle income countries, and 16 percent in 1981 to 7 percent in 1993 for low-income countries (Megginson and Netter 2011).

Proponents of privatization posit at least three ways in which it can raise welfare. First, by formally establishing property rights and making owners and managers accountable for profits and losses, the reallocation of assets from the public to the private sector can increase the operating efficiency and financial performance of firms previously owned by the state. Second, if privatization also induces entry and creates more competition, it can increase consumer surplus and the overall quality of goods and services. Third, for a given level of tax revenue, selling loss-making enterprises reduces the size of the government's deficit, frees up resources for investment in public goods, and generates revenues that can be used to pay down debt.

In the case of Latin America, fiscal constraints were a driving factor behind privatization. For years prior to Mexico's prominent debt default in August 1982, loss-making, state-owned enterprises in countries across Latin America contributed to chronic budget deficits that were the root cause of the region's debt and inflation crises. The easing of restrictions on foreign ownership of domestic equity in the late 1980s and early 1990s facilitated the stock market sale of state-owned enterprises that were a drain on public finances. For example, shares of YPF, the Argentine national oil company, were divested on the New York Stock Exchange in 1993, and Brazil conducted equity sales in electricity, steel, and telecoms in 1997.

Because privatization is generally implemented at the level of the firm, evaluating its impact on economic growth is necessarily nuanced. One exception, however, was the economies of Eastern Europe and the former Soviet Union where the massive scale of the privatization efforts effectively amounted to an aggregate shock. Given the size and scope of the shift from state to market production in these countries following the fall of the Berlin Wall, they provide an ideal setting in which to evaluate whether privatization generates aggregate efficiency gains.

In the twelve transition economies of Central and Eastern Europe and the Commonwealth of Independent States that replaced the Soviet Union, the average private-sector share of GDP rose from 13 percent in 1990 to 65 percent in 1998. These initial years of privatization were accompanied by deep recessions in Poland, Hungary, Romania, the Czech Republic, Slovakia, and the former Soviet Union due in substantial part to the massive disruptions that ensued during the transition from state to market (Blanchard and Kremer 1997; Estrin et al. 2009; Fischer 1992). In 1991, for example, the economies of Poland, Romania, and the Czech Republic contracted by an average of 10.5 percent. Between 1998 and 2007, however, these countries were also among the fastest growing economies in Europe. The average annual growth rates during this period were 3.7 percent in the Czech Republic and Hungary, 4.3 percent in Poland, 4.4 percent in Romania, 5.9 percent in the Russian Federation, and 5 percent in Slovakia.

By and large, the macroeconomic evidence suggests that post-communist privatization efforts, when accompanied by complementary reforms, may have had a positive effect on the long-run level of aggregate output (Svenjar 2002). The effects vary, however, in accordance with: 1) the speed of implementation (shock therapy versus gradualism); 2) whether ownership was subsequently dispersed or concentrated; and 3) whether the new owners of the enterprise were foreigners or domestic residents. Sale to foreign owners primarily led to positive effects on the level of total factor productivity, firm revenues, labor productivity, employment, and wages; sales to domestic residents, on the other hand, resulted in weaker or categorically negative effects (Estrin et al. 2009).

There are many reasons for the varied record of privatization across space and time. The extent to which privatization is expected to raise efficiency is complicated, subtle, and context-specific. The design of privatization programs appears paramount in putting into place the foundation for subsequent economic progress, and the mode of privatization therefore matters. Rapid privatization in Russia—especially of state-owned enterprises in oil, natural gas, and minerals—generally led to inefficiencies and corruption (Black, Kraakman, and Tarassova 2000). Gradual divestment in Poland and Slovenia was more positive (Svenjar 2002). Corporate governance and institutional frameworks are also important determinants of whether the transfer of ownership to private hands and later restructurings delivered the desired productivity gains.

Turning from macro to micro data, and moving beyond Eastern Europe and the former Soviet Union, reveals more definitive results. There are a range of studies of the financial and operating performance of firms before and after privatization that employ comprehensive data on manufacturing firms in Africa, Asia, and Latin America in addition to the transition economies. Early scholarship in the area found that real sales, operating efficiency, profitability, capital spending, and dividend payments all show significant increases, along with declining leverage (Megginson and Netter 2001; Boubakri and Cossett 1998). More recent work documents that improvements in operating performance exhibit sector- and region-specific heterogeneity. Bank performance, for example, improves significantly, but the gains from

privatization in electricity and water are limited, and the impact in telecommunications varies by region (Estrin and Pelletier 2018). Importantly, post-privatization improvements in profitability do not always result in layoffs, as a significant fraction of privatized firms actually employed more workers (Megginson and Netter 2001; Gupta 2005; Estrin et. al. 2009). The caveat in all of this, of course, is that if governments systematically privatize firms that are already better-positioned in some way, selection bias may lead to erroneous attribution of improved performance to the merits of private ownership (Dinc and Gupta 2011; Gupta, Ham, and Svejnar 2008).

Another concern about privatization stems from evidence in a sample of privatized firms in 39 emerging and developed countries that ownership becomes more concentrated in the two decades following divestment (Boubakri, Cosset, and Guedhami 2005). The risk and reality of increased concentration demonstrates that improved operating performance does not necessarily imply positive-sum outcomes. Indeed, given the rents generated in some cases for the lucky few who were able to acquire state assets, significant controversy surrounds the question of increased market power rather than broad-based welfare gains following privatization.

Concerns about levels of rents and ownership concentration were flagged early on during privatization efforts in Russia and Eastern Europe. Prominent examples include small groups of oligarchs who managed to concentrate power quickly and accumulate wealth, tainting the reputations of privatization programs through indictments of corruption and cronyism (Roland 2008). Measures to minimize concentration included calls to incentivize "divesture commissions" to perform breakups in industries where there were concerns about anti competitive behavior, and recommendations to perform market structure interventions to prevent collusion before divestiture (Tirole 1991). Privatization critics argue that neither public nor private provision can fully resolve the difficult incentive problems and the choice simply depends on the transaction costs associated with future public or private interventions (Sappington and Stiglitz 1987). While the benefits of privatizing competitive industries are less controversial, on balance, state-owned natural monopolies may be preferable if they mitigate regressive redistributive effects.

Finally, an underappreciated nuance of ownership concentration is that whether under state or private control, ownership concentration and regulatory capture can delay or stall other reforms such as the liberalization of foreign direct investment. Evidence suggests that the propensity to open up industries to foreign investment is inversely related to industry concentration (Chari and Gupta 2008). Efficiency gains are compromised when reform movements are highjacked by special interests, which suggests that the political economy of privatization has significant implications for efficiency.

Resistance and Resentment

Efficiency-enhancing policy changes often involve difficult adjustments. In democratic settings, enough of those who might form a coalition blocking such

changes must be persuaded to back them (or at least not to oppose them actively) if the reform process is to be sustained (Brady and Spence 2009). Therefore, in addition to strategic knowledge about the aggregate benefits that such policy changes could bring, successful reform requires tactical knowledge—and bargaining chips. As Krueger (1988) wrote: "[T]he economic policies that lead to debt difficulties (and those that lead to rapid growth) are intensely political . . . . [T]he international community has thus far failed to find techniques to reward adherence to altered policy packages over a sustained period . . . If one were to identify one desirable type of financing facility, it would . . . simultaneously increase the credibility of the program, serve as an additional inducement to undertake appropriate reform measures, and overcome debt overhang."

We have demonstrated that the Baker Hypothesis for how to improve growth stands up quite well to empirical scrutiny. The same cannot be said for Baker's (1985) official three-step plan for bringing about reform. The Baker Plan stumbled, in large part, because it rejected Krueger's point about debt overhang, which is the situation in which a country rules out additional borrowing, even for worthwhile purposes, because its current debts are so high that all the benefits of new borrowing would accrue to existing debtholders (Krugman 1988; Sachs 1989). Under the first step of the Baker Plan, leaders had to implement reforms to maintain access to official lending from the IMF and World Bank. Second, their countries would start growing as a result of the first step. Third, private creditors (the commercial banks) would voluntarily resume lending because of the second step. The Baker Plan did not ask the banks to write down debt to eliminate overhang, nor did it hold banks at least somewhat accountable for extending ill-advised loans in the past. Baker explicitly and publicly opposed any form of debt relief (as discussed in Arslanalp and Henry 2005).

Some changes did take place in the immediate aftermath of Baker's speech (Williamson 2004). Colombia and Costa Rica, for example, liberalized trade in 1986, and a number of debt-laden countries undertook minimal reforms to retain access to IMF and World Bank money. But without debt relief, not enough leaders had the political capital they needed to drive sustained economic transformation. The reform dates discussed earlier illustrate the point. The average stabilization year was 1992—seven years after Baker's speech—and the average trade liberalization year was 1990.

In order to accelerate the reform process, Nicholas F. Brady, Baker's successor at the US Treasury, announced a new financing facility in May 1989. In return for countries agreeing to implement and sustain the kinds of economic policy changes emphasized by Baker, countries were offered debt relief that would eliminate debt overhang and clear the way for new, profitable private lending. Once countries managed to negotiate a debt-reduction agreement, their implementation of reforms under the Brady Plan was swift. For the 16 countries that eventually received debt relief, the average year of reaching a Brady agreement was 1992—the same as their average stabilization year and two years after their average trade liberalization date.

The accomplishments of the Brady Plan notwithstanding, Baker's uneven treatment of the debt overhang problem had lasting ramifications. Specifically, Baker's insistence that economic restructuring take place without the banks' accepting meaningful responsibility left the leaders of many developing countries in a politically untenable position and ignited a firestorm of criticisms from multiple sources that were united by a theme of enduring resonance: that is, the theme that Washington, Wall Street, and the leaders of the international financial system resolved a banking crisis by driving through policy changes that hurt the common man and helped the bankers.

The consequences of failing to address debt overhang in the 1980s appears to have had some impact on the IMF and the World Bank, as they gradually adopted a more flexible approach to reforms. For instance, when the Asian financial crisis hit in late 1997, the IMF initially insisted on fiscal austerity, but changed tack in April 1998, allowing crisis-impacted governments to swing, on average, from a primary surplus of 1.8 percent of GDP to an average deficit of 1.8 percent (Chari and Henry 2015). The IMF displayed similar flexibility during the eurozone debt crisis, playing a central role in the government of Greece securing debt relief from its creditors, even though the arrangements were not finalized until June 2018. Debt overhang arguments also figured prominently in debt relief initiatives for the world's poorest countries, such as the 2005 Gleneagles Declaration and the World Bank and IMF's 2020 Debt Service Suspension Initiative in response to the pandemic.[8]

At present, leaders around the globe are grappling with the COVID-19 pandemic, along with increasing discontent over rising inequality and fears about the environment and climate change. But these legitimate concerns should not eclipse one of the most important stories about the world economy since the end of World War II, even as it continues to unfold. Certain economic policy reforms implemented by emerging and developing economies have significantly improved their economic performance, helping to lift hundreds of millions of people out of poverty with positive attendant consequences for health and life expectancy. Macroeconomic stability and economic efficiency are not sufficient conditions for a flourishing society, but they are absolutely necessary for sustainable and inclusive growth that allows an increasing fraction of a country's population to have choices and opportunity. The hard-won economic successes of the past three decades underscore the benefits of policymakers finding the will and the ways to meaningfully and constructively address the prospect of continued catch-up growth by emerging and developing economies.

[8]Arslanalp and Henry (2006) demonstrate that the debt situations of the world's poorest countries are sui generis and that debt overhang logic does therefore not apply to them.

■ *We are grateful to Nandini Gupta, Paul Romer, Alejandro Werner, and seminar participants at the Hoover Institution and UC Santa Cruz for helpful comments, and to Sandile Hlatshwayo and Bill Megginson for help with the privatization data. Henry and Reyes thank the Sloan Foundation for financial support of the PhD Excellence Initiative, of which Reyes is Fellow and Henry is Principal Investigator. Henry is a member of the Board of Directors of Citigroup and Nike; neither corporation has funded this research.*

References

Amsden, A. H. 1989. *Asia's Next Giant: South Korea and Late Industrialization*. New York: Oxford University Press.

Andrés, Javier, and Ignacio Hernando. 1999. "Does Inflation Harm Economic Growth? Evidence from the OECD." In *The Costs and Benefits of Price Stability*, edited by Martin Feldstein. Chicago: The University of Chicago Press.

Arslanalp, Serkan, and Peter Blair Henry. 2005. "Is Debt Relief Efficient?" *Journal of Finance* 60 (2): 1017–51.

Arslanalp, Serkan, and Peter Blair Henry. 2006. "Debt Relief." *Journal of Economic Perspectives* 20 (1): 207–20.

Baker, James A. III. 1985. "Statement of the Honorable James A. Baker III, Secretary of the Treasury of the United States Before the Joint Annual Meeting of the International Monetary Fund and the World Bank, October 8, 1985, Seoul, Korea" in *International Bank for Reconstruction and Development, International Finance Corporation, International Development Association. Annual Meetings of the Board of Governors Summary Proceedings*.

Balassa, Bela. 1977. *Policy Reform in Developing Countries*. Oxford: Pergamon Press.

Black, Bernard, Reinier Kraakman, and Anna Tarassova. 2000. "Russian Privatization and Corporate Governance: What Went Wrong?" *Stanford Law Review* 52 (6): 1731–1808.

Blanchard, Olivier, and Michael Kremer. 1997. "Disorganization." *The Quarterly Journal of Economics* 112 (4): 1091–1126.

Boubakri, Narjess, and Jean-Claude Cosset. 1998. "The Financial and Operating Performance of Newly Privatized Firms: Evidence from Developing Countries." *The Journal of Finance* 53 (3): 1081–1110.

Boubraki, Narjess, Jean-Claude Cosset, and Omrane Guedhami. 2005. "Postprivatization Corporate Governance: The Role of Ownership Structure and Investor Protection." *Journal of Financial Economics* 76 (2): 369–99.

Brady, David, and Michael Spence. 2009. "Leadership and Politics: A Perspective from the Growth Commission." *Oxford Review of Economic Policy* 25 (2): 205–18.

Bruno, Michael, and William Easterly. 1998. "Inflation Crises and Long-Run Growth." *Journal of Monetary Economics* 41 (1): 3–26.

Bulow, Jeremy. 2002. "First World Governments and Third World Debt." *Brookings Papers on Economic Activity* 1: 229–55.

Chari, Anusha, and Peter Blair Henry. 2004. "Risk Sharing and Asset Prices: Evidence from a Natural Experiment." *Journal of Finance* 59 (3): 1295–1324.

Chari, Anusha, and Peter Blair Henry. 2008. "Firm Specific Information and the Efficiency of Investment." *Journal of Financial Economics* 87 (3): 636–55.

Chari, Anusha, and Nandini Gupta. 2008. "Incumbents and Protectionism: The Political Economy of Foreign Entry Liberalization." *Journal of Financial Economics* 88 (3): 633–56.

Chari, Anusha, and Peter Blair Henry. 2015. "Two Tales of Adjustment: East Asian Lessons for European Growth." *IMF Economic Review* 63: 164–96.

Chari, Anusha, Peter Blair Henry, and Diego Sasson. 2012. "Capital Market Integration and Wages." *American Economic Journal: Macroeconomics* 4 (2): 101–32.

Chomsky, Noam. 1999. *Profit Over People: Neoliberalism and Global Order.* New York: Seven Stories Press.

Commission on Growth and Development. 2008. *The Growth Report: Strategies for Sustained Growth and Inclusive Development.* Washington, DC: World Bank.

Coulibaly, Brahima, Belinda Archibong, Ngozi Okonjo-Iweala. 2021. "Washington Consensus Reforms and Economic Performance in Sub-Saharan Africa." AGI Working Paper 27.

Dinc, I. Serdar, and Nandini Gupta. 2011. "The Decision to Privatize: Finance and Politics." *Journal of Finance* 66 (1): 241–69.

Dornbusch, Rüdiger, and Stanley Fischer. 1993. "Moderate Inflation." *World Bank Economic Review* 7 (1): 1–44.

Dornbusch, Rüdiger, and Yung Chul Park. 1987. "Korean Growth Policy." *Brookings Papers on Economic Activity* 2: 389–454.

Easterly, William. 1996. "When Is Stabilization Expansionary? Evidence from High Inflation." *Economic Policy* 11 (22): 65–107.

Easterly, William. 2019. "In Search of Reforms for Growth: New Stylized Facts on Policy and Growth Outcomes." NBER Working Paper 26318.

Edwards, Sebastian. 1995. *Crisis and Reform in Latin America: From Despair to Hope.* New York: Oxford University Press.

Estevadeordal, Antoni, and Alan Taylor. 2013. "Is the Washington Consensus Dead? Growth, Openness, and the Great Liberalization, 1970s–2000s." *Review of Economics and Statistics* 95 (5): 1669–90.

Estrin, Saul, Jan Hanousek, Evzen Kocenda, and Jan Svejnar. 2009. "The Effects of Privatization and Ownership in Transition Economies." *Journal of Economic Literature* 47 (3): 699–728.

Estrin, Saul, and Adeline Pelletier. 2018. "Privatization in Developing Countries: What Are the Lessons of Recent Experience?" *World Bank Research Observer* 33 (1): 65–102.

European Bank for Reconstruction and Development. 2007. *Transition Report 2007: People in Transition.* London: European Bank for Reconstruction and Development

Feldstein, Martin. 2002. "Economic and Financial Crisis in Emerging Market Economies: Overview of Prevention and Management." NBER Working Paper 8837.

Fischer, Stanley. 1986. "Issues in Medium-Term Macroeconomic Adjustment." *World Bank Observer* 1 (2): 163–82.

Fischer, Stanley. 1987. "Economic Growth and Economic Policy." In *Growth-Oriented Adjustment Programs*, edited by Vittorio Corbo, Morris Goldstein, and Mohsin Khan, 151–178. Washington, DC: IMF and The World Bank.

Fischer, Stanley. 1992. "Stabilization and Economic Reform in Russia." *Brookings Papers on Economic Activity* 1: 77–126.

Fischer, Stanley. 1993. "The Role of Macroeconomic Factors in Growth." *Journal of Monetary Economics* 32 (3): 485–512.

Fischer, Stanley. 2003. "Globalization and Its Challenges." *American Economic Review* 93 (2): 1–30.

Gertz, Geoffrey, and Homi Kharas 2019. *Beyond Neoliberalism: Insights from Emerging Markets.* Brookings Institution Report. Washington, DC: Global Economy and Development at Brookings.

Gil Diaz, Francisco. 2003. "Don't Blame Our failures on Reforms That Have Not Taken Place." *Fraser Forum* : 7–10.

Global Financial Data Database. 2021. "GFD Indices." *GFD Finaeon.* https://finaeon-globalfinancialdata-com.proxy.library.nyu.edu/ (accessed June 28, 2021).

Grier, Kevin B., and Robin M. Grier. 2021. "The Washington Consensus Works: Causal Effects of Reform, 1970–2015." *Journal of Comparative Economics* 49 (1): 59–72.

Gupta, Nandini. 2005. "Partial Privatization and Firm Performance." *Journal of Finance* 60 (2): 987–1015.

Gupta, Nandini, Jhon C. Ham, and Jan Svejnar. 2008. "Priorities and Sequencing in Privatization: Evidence from Czech Firm Panel Data." *European Economic Review* 52 (2): 183–208.

Ha, Jongrim, M. Ayhan Kose, and Franziska Ohnsorge. 2019. *Inflation in Emerging and Developing Economies: Evolution, Drivers, and Policies.* Washington, DC: The World Bank.

Henry, Peter Blair. 2000a. "Stock Market Liberalization, Economic Reform, and Emerging Market Equity Prices." *Journal of Finance* 55 (2): 529–64.

Henry, Peter Blair 2000b. "Do Stock Market Liberalizations Cause Investment Booms?" *Journal of Financial Economics* 58 (1–2): 301–34.

Henry, Peter Blair. 2002. "Is Disinflation Good for the Stock Market?" *Journal of Finance* 57 (4): 1617–48.

Henry, Peter Blair. 2007. "Capital Account Liberalization: Theory, Evidence, and Speculation." *Journal of Economic Literature* 45 (4): 887–935.

Henry, Peter Blair. 2013. *Turnaround: Third World Lessons for First World Growth*. New York: Basic Books.

International Monetary Fund. 2001. "The Decline of Inflation in Emerging Markets: Can It Be Maintained?" In *World Economic Outlook*. Washington, DC: International Monetary Fund.

International Monetary Fund Database. 2021. "Access to Macroeconomic and Financial Data." *International Financial Statistics Database*. https://data.imf.org/?sk=4c514d48-b6ba-49ed-8ab9-52b0c1a0179b&sId=1409151240976 (accessed June 28, 2021).

Irwin, Douglas A. 2019. "Does Trade Reform Promote Economic Growth? A Review of Recent Evidence." NBER Working Paper 25927.

Jao, Y.C. 1976. "Trade and Economic Development in Taiwan." *Intereconomics* 11 (6): 172–76.

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

Krueger, Anne, 1988. "Resolving the Debt Crisis and Restoring Developing Countries Creditworthiness." Paper presented at the Carnegie-Rochester Public Policy Conference, University of Rochester.

Krueger, Anne O. 1995. "East Asian Growth Experience and Endogenous Growth Theory." In *Growth Theories in Light of the East Asian Experience*, edited by Takatoshi Ito and Anne O. Krueger. Chicago: University of Chicago Press.

Krueger, Anne. 2004. "Meant Well, Tried Little, Failed Much: Policy Reforms in Emerging Market Economies." Roundtable lecture presented at the Economics Honor Society, New York University, March 23.

Krugman, Paul. 1988. "Financing versus Forgiving a Debt Overhang." *Journal of Development Economics* 29 (30): 253–68.

Krugman, Paul. 1994. "Does Third World Growth Hurt First World Prosperity." *Harvard Business Review,* July–August 1994. https://hbr.org/1994/07/does-third-world-growth-hurt-first-world-prosperity.

Lewis, W. Arthur 1954. "Economic Development with Unlimited Supplies of Labour." *The Manchester School* 22 (2): 139–91.

Mathieson Donald J., and R. McKinnon. 1981. "How to Manage a Repressed Economy." *Essays in International Finance* 145.

McKinnon, Ronald I. 1984. "The International Capital Market and Economic Liberalization in LDCs." *Developing Economies* 22 (4): 476–81.

Megginson, William L., and Jeffry M. Netter. 2001. "From State to Market: A Survey of Empirical Studies on Privatization." *Journal of Economic Literature* 39 (2): 321–89.

Menon, Ravi. 2015. "An Economic History of Singapore: 1965–2065." Keynote address delivered at the Singapore Economic Review Conference, Singapore, August 5.

Michalopolous, Constantine. 1987. "World Bank Programs for Adjustment and Growth." World Bank Discussion Paper ERS 11.

Obstfeld, Maurice. 1998. "The Global Capital Market: Benefactor or Menace?" *Journal of Economic Perspectives* 12 (4): 9–30.

Rodríguez, Francisco, and Dani Rodrik. 2000. "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence." *NBER Macroeconomics Annual* 15 (1): 261–325.

Rodrik, Dani. 2006. "Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's *Economic Growth in the 1990s: Learning from a Decade of Reform*." *Journal of Economic Literature* 44 (4): 973–87.

Rogoff, Kenneth. 1999. "International Institutions for Reducing Global Financial Instability." *Journal of Economic Perspectives* 13 (4): 21–42.

Roland, Gérard, ed. 2008. *Privatization: Successes and Failures*. New York: Columbia University Press.

Sachs, Jeffrey. 1987. "Trade and Exchange Rate Policies in Growth-Oriented Adjustment Programs." In *Growth-Oriented Adjustment Programs*, edited by Vittorio Corbo, Morris Goldstein, and Mohsin Khan. Washington, DC: IMF and The World Bank.

Sachs, Jeffrey. 1988. "Conditionality, Debt Relief and the Developing Country Debt Crisis." NBER Working Paper 2644.

Sachs, Jeffrey. 1989. "The Debt Overhang of Developing Countries" In *Debt, Stabilization and Development*, edited by Guillermo A. Calvo, Ronald Findlay, Pentti Kouri, and Jorge Braga De Macedo. Oxford: Basil Blackwell, Oxford.

Sachs, Jeffrey D., and Andrew Warner. 1995. "Economic Reform and the Process of Global Integration." *Brookings Papers on Economic Activity* 1: 1–118.

Sappington, David, and Joseph Stiglitz. 1987. "Privatization, Information and Incentives." *Journal of Policy Analysis and Management* 6 (4): 567–85.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70 (1): 65–94.

Stiglitz, Joseph 2002. *Globalization and Its Discontents.* New York: W.W. Norton & Company.

Stulz, René M. 1999. "Globalization of Equity Markets and the Cost of Capital." NBER Working Paper 7021.

Svejnar, Jan. 2002. "Transition Economies: Performance and Challenges." *Journal of Economic Perspectives* 16 (1): 3–28.

Tirole, Jean. 1991. "Privatization in Eastern Europe: Incentives and the Economics of Transition." *NBER Macroeconomics Annual* 6: 221–68.

Wacziarg, Romain, and Karen Welch. 2008. "Trade Liberalization and Growth: New Evidence." *World Bank Economic Review* 22 (2): 187–231

Williamson, John. 1990. "What Washington Means by Policy Reform." *In Latin American Adjustment: How Much Has Happened?* edited by John Williamson. Washington, DC: Institute for International Economics.

Williamson, John. 2004. "A Short History of the Washington Consensus." https://www.piie.com/publications/papers/williamson0904-2.pdf.

World Bank Database. 1960–2020. "Inflation, consumer prices (annual %)." International Monetary Fund international financial statistics and data files. https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG (accessed June 28, 2021).

World Bank Database. 1961-2019. "GDP growth (annual %)." World Bank national accounts data and OECD National Accounts data files. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG (accessed June 28, 2021).

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 109–132

Washington Consensus in Latin America: From Raw Model to Straw Man

Ilan Goldfajn, Lorenza Martínez, and Rodrigo O. Valdés

The Washington Consensus emerged in the second half of the 1980s after a prolonged period of poor economic performance in Latin America, initially reflecting a necessity more than a thoughtfully conceived plan. A number of countries in Latin America had previously borrowed heavily in US dollars, and the era of tight monetary policy under US Federal Reserve chair Paul Volcker left them in the doldrums. Several countries experienced debt defaults (from Mexico in 1982 to Brazil in 1987), deep recessions, and banking crises. In Latin America, the 1980s became known as the Lost Decade.

Latin America's debt crisis and the associated problems led to the need to reestablish financing by the private sector. US Treasury Secretary James Baker (1985) outlined a "Program for Sustained Growth" for these countries at a joint meeting of the IMF and the World Bank. A central element was that the debtor countries adopt market-oriented policies to create "more flexible and productive economies" (p. 209). John Williamson (1990b) initially coined the term Washington Consensus at a conference organized in 1989 to acknowledge the ongoing efforts made by Latin American countries in implementing structural reforms in line with macroeconomic prudence, trade liberalization, opening to foreign direct investment, and privatization, among other structural reforms that would

■ *Ilan Goldfajn is Founder and Director of Center for Public Policy Debates and Chairman of Credit Suisse, Sao Paulo, Brazil. Lorenza Martínez is Executive Director of Actinver, Mexico City, Mexico. Rodrigo Valdés is Associate Professor, School of Government, Pontificia Universidad Católica de Chile, Santiago, Chile. Their email addresses are Ilan@cdpp.org.br, lorenzamtzt@gmail.com, and rodrigo.valdes@uc.cl.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.109.

attract private capital through higher expected potential output growth. Different IMF/World Bank–supported programs in the region also became blueprints of the plan.

The overarching expectation of governments, economic analysts, and financial markets alike was that the reforms would reestablish macroeconomic stability and prompt renewed growth. The Mexican economist Pedro Aspe (1993, pp. 46–47), who served as Minister of Finance in Mexico from 1988 to 1994, summarized the attitude: "The economic strategy based on fiscal and monetary discipline, consensus gathering, and the reform of the state has already yielded very encouraging results, not only in terms of short-term macroeconomic performance, but also in creating new prospects for sounder long-term growth." If one looks back at the five-year GDP growth forecasts included in the different IMF *World Economic Outlook* vintages, the forecasts for the early 1990s often suggested annual growth rates of 5 or 6 percent for most Latin American countries (as shown in the online Appendix available with this paper at the *JEP* website).

But controversies over the Washington Consensus immediately blossomed. Across Latin America, several political groups opposed the Washington Consensus policies, for two main reasons: some saw them as imposed by the United States in an effort to increase its control over Latin American countries and promote the interests of international companies; while others considered that these policies had already been tried in the 1980s and had failed to stabilize the economy, entailing a high economic cost.

Even at the outset, it was unclear that these were "consensus" policies. Williamson's (1990a) ten-point list was a descriptive exercise of what was happening in the context of the debt crisis, and it thus tended to focus on areas that were already being covered in early reformer countries like Chile, Mexico, and Bolivia. A number of analysts argued for a more prescriptive list that would identify other topics, including environmental policies and a clear plan to fight poverty. Furthermore, most countries implemented only parts of the reforms, and results were mixed. After decades of disappointing economic growth, the five-year GDP growth forecasts from the IMF have now declined significantly for all Latin American countries, reaching a meager 2–3 percent even before the pandemic recession.

In this paper, we begin with Williamson's (1990) ten-point Washington Consensus and explore how Latin American countries responded, or didn't, to the recommendations. We then present short case studies of Brazil, Mexico, and Chile. Brazil and Mexico are chosen because together they account for more than half of the total population and GDP in the region as a whole. Chile, in turn, was commonly viewed as the poster child for economic reform in Latin America in the 1980s, but the reforms were implemented by a military dictatorship until democratic elections returned in 1989. An important question in the region, then, was whether these initially painful reforms were doable in a democracy. Moreover, these three countries are each a home base for one of the authors of this paper. We emphasize that despite some broad similarities, the degree of implementation and the timing of the Washington Consensus policies varied substantially across countries.

For varying reasons, all three countries would become dissatisfied with the direction of economic policy over time.

Following the case studies, we assess the performance of the Washington Consensus in hindsight, after 30 years, considering metrics like inflation, productivity, and growth. We also look at poverty and inequality because they are relevant indicators of welfare—although they were not directly the focus of the Washington Consensus reforms. A fair assessment of the performance needs to recognize that no country adopted the Washington Consensus exactly as it was designed or as it was implemented in other countries. Relative success depended not only on the degree of implementation but also on country specificities and external shocks. In addition, we draw a distinction between core Washington Consensus policies that were enacted in the late 1980s and early 1990s and new policies that were implemented in the late 1990s and the 21st century. It is important to acknowledge that today's lens on the Washington Consensus policies differs substantially from the original perspective, because the situation, the stock of knowledge, and social values have all changed over time.

What Was Adopted? The Reality of the Washington Consensus in Latin America

Following John Williamson's (1990) ten overarching principles for the Washington Consensus, we offer here an overview of how they were implemented in Latin America.

1. Fiscal discipline, with a deficit of 1 to 2 percentage points of GDP considered adequate.

Most Latin American countries did not achieve the goal of a fiscal deficit below 2 percent of GDP on a sustained basis. Initially, many countries in the region—including Chile and Mexico—achieved significant progress with a combination of contained deficits and growth (often tied to IMF programs). Later, however, results became deeply heterogeneous. A few countries like Chile, Colombia, Peru, and Mexico managed to contain deficits and debt despite the Asian Crisis of 1997, which worsened external financial conditions markedly. Others entered complex dynamics requiring new IMF programs, and a few ended in default, like Argentina and Ecuador around 2000. In the following decade, a few countries, notably Chile and Peru, continued lowering debt significantly, partly thanks to very high export prices—the so-called "commodity price super-cycle." Others kept debt at manageable levels like Colombia and Mexico, while a few continued to be marked by fiscal challenges. After important fiscal consolidation in the late 1990s and early 2000s, Brazil continues to face fiscal challenges and currently has a higher debt than its peers. The region again saw sovereign defaults in Ecuador (2008), Argentina (2014), and Venezuela (2017).

Within fiscal measures, pension reform was another notable policy shift, although it was not directly linked to the original Washington Consensus. Several

countries reformed their old-age pension schemes into a fully funded system, including Chile in the 1980s and Argentina, Colombia, Mexico, and Peru in the 1990s. These reforms became a way to escape the medium-run fiscal pressure from pay-as-you-go systems as well as a powerful tool for developing the capital market.

2. Public expenditure reallocation into priority sectors, namely, education, health, and public investment.

In public discussions of the Washington Consensus, people are sometimes surprised to discover that reallocation of public expenditures into priority sectors was its second point. Indeed, a common criticism of the Washington Consensus is that it paid insufficient attention to education and health. Of course, this recommendation to reallocate spending into priority sectors probably reflected concerns about productivity growth, rather than direct social and anti-poverty efforts.

It seems fair to say that the original Washington Consensus policies largely neglected income distribution and other social issues (such as social mobility) and never consolidated them in an organized way. Similarly, there was a lack of emphasis on education as an essential social mobility tool and a key ingredient of long-run growth. However, since the second half of the 1990s, these issues have become an increasingly important part of the agenda, and Latin America has seen an increase in spending on social programs. In Brazil, expenditures on social programs (such as conditional cash transfers to the poor through Bolsa Familia) increased from 9.8 percent of total spending in 1997 to 17.4 percent in 2019. In Mexico, social expenditure, including education, health, and poverty alleviation programs, increased from 30 percent of total public spending in the 1980s to 51 percent in the 1990s and 68 percent in the 2010s. In the same period, the expenditure share of previously state-owned firms and public investment declined. In Chile, between the 1990s and the 2010s, the share of education and health in total spending increased from 25 percent to 40 percent; this was made possible by cuts in defense and pensions (due to the end of the pay-as-you-go system).

3. Tax schemes characterized by a broad tax base, moderate marginal tax rates, and a strong tax administration, as fiscal revenues had to support the needed public investment and expenditure.

Countries across Latin America cut their top tax rates (Lora 2001; Trading Economics 2021; and OECD Tax Database 2021). Between 1986 and 1999, the median maximum personal income tax rate was slashed by 20 percentage points and the top corporate tax rate by 8 percentage points. The maximum personal income tax rate was cut from 60 to 25 percent in Brazil, from 55 to 35 percent in Mexico, and from 50 to 45 percent in Chile. In that same time frame of 1986 to 1999, the top corporate income tax rate fell from 45 to 25 percent in Brazil and from 42 to 34 percent in Mexico—although it rose slightly from 10 to 15 percent in Chile. In all three countries, the value-added tax rate remained relatively stable in the range of 15 to 20 percent, although Mexico maintained reduced value-added tax rates for specific regions and certain goods.

Taking these changes as a whole, the share of consumption tax revenues declined somewhat in Brazil, Mexico, and Chile, while the share of income taxes (the sum of corporate and personal) grew modestly. In Brazil, for example, consumption tax revenues fell from 48 percent of total tax revenue in the 1990s to 46 percent in the 2000s, and income tax revenues rose from 19 to 21 percent of all revenues during that span.

In general, with the exception of Argentina and Brazil, income tax revenues and total tax revenues have remained low in Latin American compared to higher-income countries. For example, total tax revenues increased significantly in Mexico after the 1994 "tequila crisis," but remain well below 20 percent of GDP. Total tax revenues in Chile barely rose to 19 percent. In Brazil, significant indirect taxes and other types of revenue dominated, elevating the total tax burden to 32–35 percent of GDP.

4. Market-determined interest rates and real rates at moderate positive (or at least not negative) levels.

Practically all Latin American countries liberalized interest rates between 1985 and 2000 (Lora 2001). By 1992, all countries in South America had freed interest rates. Although some countries maintained some earmarked lending, the region moved toward global banking standards relatively quickly. Since prudential financial regulation was strengthened and Basel regulatory standards were adopted, there have not been any widespread banking troubles. In Brazil, Mexico, and Chile, the financial sector has been quite resilient despite occasional large shocks. Of course, the region has seen other banking crises—for example, Argentina in 2001–02 and Ecuador in 1998—but it is difficult to connect those to the Washington Consensus; rather, they were part of macroeconomic experiments that went wrong.

During these 30 years, the Latin American financial system deepened significantly, and financial liberalization contributed to both access to financing, especially in nontradable sectors, and economic growth (Tornell, Westermann, and Martínez 2003). Progress in liberalizing financial markets is also reflected in global market access and foreigners' participation in local debt markets, which have developed substantially—also fostered by private pension savings (Borensztein et al. 2008).

5. Competitive exchange rates to support export-led growth, while avoiding multiple exchange-rate regimes, where the exchange rate could either be market-determined or set at a level consistent with a sustainable current account deficit.

Exchange rate regimes in Latin America generally became more flexible in the 1990s. However, "intermediate" exchange-rate regimes (in the middle ground between floating and fixed) were still prevalent, which allowed for some but not full flexibility. Crawling exchange-rate bands and pegs that were adjusted only occasionally were subject to speculative attacks. Since the late 1990s, countries in the region have moved away from such intermediate exchange-rate regimes, because they discouraged firms and investors from managing exchange-rate risk and thus could lead to periods of false stability punctuated by disruptive shocks.

In the last decade, the majority of Latin American countries maintained floating exchange rate systems, although still leaving open the possibility of occasional intervention in special circumstances. Overall, markets have clearly had an increasing role in determining the exchange rate (Levy-Yeyati and Sturzenegger 2016; Ilzetzki, Reinhart, and Rogoff 2019). Multiple official and unofficial exchange rates are a thing of the past, except in Argentina and Venezuela.

The monetary and exchange-rate framework in Latin America has gone well beyond what was initially envisaged by the Washington Consensus. With inflation targeting by central banks serving to anchor price levels, authorities have become more comfortable with allowing the exchange rate to act as a buffer for shocks. Moreover, despite a common belief, the Washington Consensus did not call for the removal of capital controls as a priority, because this policy lacked consensus among economists and policymakers at the time. Nonetheless, many countries in Latin America have eliminated their historical capital controls, including Brazil, Mexico, and Chile.

6. Trade policy aimed at liberalizing imports to allow exporters access to the necessary capital and intermediate goods to be competitive in international markets; in particular, reducing tariffs to 10 to 20 percent, with low variance and removing all other forms of import barriers.

Latin America has advanced toward greater openness to trade, but with some notable exceptions. Chile and Mexico (and later Peru and to some extent Colombia) opened up to trade by cutting tariffs and signing free trade agreements with crucial partners, thus embracing an open-economy development strategy. By different measures, they have become more trade-integrated than many industrialized countries.

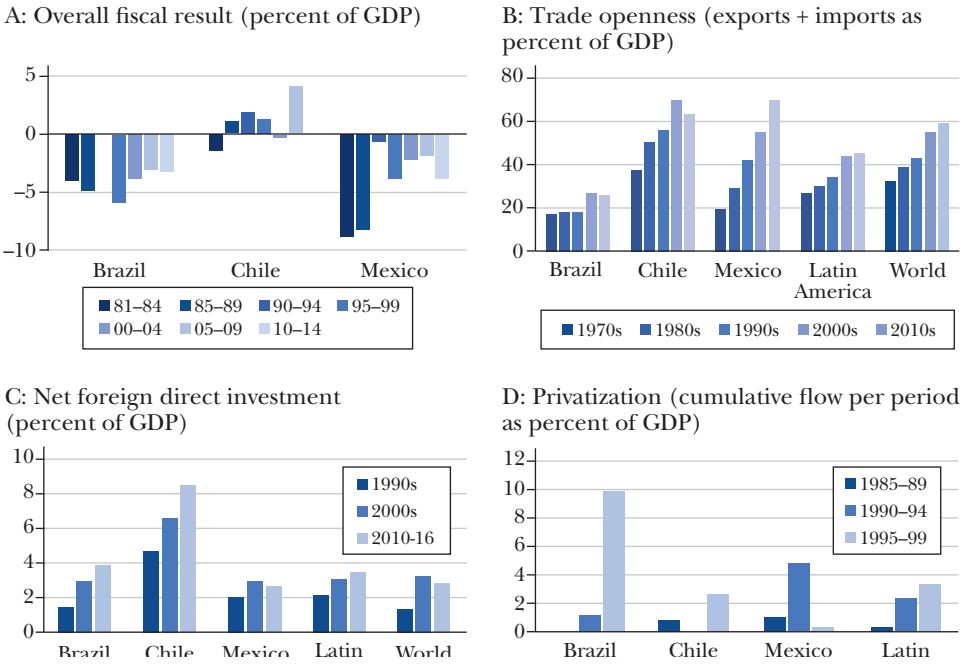
Brazil (and Argentina), in contrast, cut some tariffs but kept key import barriers. Protectionism and the idea of a growth strategy based on import substitution is still part of the ideological matrix of the private sector. In comparison with the world average, Brazil remains a closed economy (as shown in Figure 1B). Similar patterns emerge from other sources like the "de jure trade openness" measure calculated by the KOF Swiss Economic Institute Globalization Index.[1]

7. Opening to foreign direct investment as a way to obtain much-needed capital investment, along with skills and know-how.

Latin America has opened to foreign direct investment but with mixed results. While net inflows to Latin America increased (Figure 1C), the regional average barely surpassed the world average. Inflows to Brazil and Chile have more than doubled since the 1990s (including both green- and brownfield investments). In

[1] The KOF trade globalization de jure index is calculated as the weighted average of five variables: trade regulations or non-tariff trade barriers and compliance costs of importing and exporting, trade taxes calculated as the income from taxes on international trade as percentage of total revenue (inverted), the unweighted mean of tariff rates and the number of bilateral and multilateral free trade agreements.

Figure 1
Some Evidence on Implementing the Washington Consensus in Latin America



Source: Panel A: Mauro et al. 2013 and International Monetary Fund World Economic Outlook databases. For Brazil, Banco Central do Brasil, Department of Statistics (2021). Panels B and C: World Bank World Development Indicators. Panel C: World Bank. Panel D: Lora (2001).
Note: Privatization refers to the proceeds received by the government from the partial or complete sale of company shares to the private sector.

contrast, Mexico has not been able to attract significant net inflows despite having a privileged geographical position for US offshoring. There has also been an important difference between oil-rich and mining countries in the region. The former have generally decided to maintain the energy sector exclusively in the hands of the state, whereas the latter expanded private mining significantly.

8. Privatization to relieve public deficits and improve efficiency and competition.

Latin America saw an immense privatization push in the 1990s, with cumulative proceeds of 6 percent of GDP (Figure 1D). This total represents close to 60 percent of all emerging market privatization revenues in that decade (Chong and López-de-Silanes 2005). The economic share of state-owned enterprises in Latin America fell from 10 percent in the late 1980s to 5 percent by the late 1990s. This positioned the region slightly above state-owned (relative) activity in industrialized countries and well below Asia. In Brazil, Mexico, and Chile, a good part of once state-run services

is now private. Oil remains mainly state-owned. Chong and López-de-Silanes (2005) find that privatized firms' profitability and efficiency increased, closing their gap against private sector benchmarks. However, the authors also find that many privatizations were not accompanied by adequate contract design and regulation, and they suffered from regulatory capture.

9. Deregulation to promote competition by eliminating different types of barriers to entry or privileges to specific firms.

Fostering competition has been a rocky road in Latin America. Antitrust institutions have developed only gradually, and there are areas where contestability is still limited in some countries (for example, airline routes). Profitability in specific industries has been abnormally high (for example, in banking and the private pension system). The Product Market Regulation Index published by the Organization for Economic Cooperation and Development (OECD) reveals that Brazil ranks very low, Chile is average, and Mexico is in between, despite some progress in absolute levels in the last ten years.[2]

10. Strengthening of property rights, which were viewed as fundamental to the proper functioning of the economy and specifically the promotion of private investment.

Various indicators of (relative) property rights protection and the rule of law show that progress has been unimpressive and somewhat uneven. We focus here on the Political Risk Services (PRS) International Country Risk Guide, as it is the standard for growth empirics (Barro 2015), and the World Bank Governance Indicators, which has the highest correlation with changes in future growth (Díaz and Valdés 2020).[3] On average, between 1996 and 2006, South American countries plus Mexico recorded a decline in their percentile rating in the PRS Rule of Law category and then remained stable in the following decade. On the World Bank measure of Control of Corruption, the average South American country improved about 3 percentage points between 1996 and 2006 but suffered a larger setback in the following ten years.

By both measures, Chile consistently ranks higher than Brazil and Mexico. For example, in the PRS Rule of Law rating, Brazil increased from 3 percent in 1996 to 12 percent in 2006, while over that interval Mexico rose from 3 to 37 percent and Chile rose from 60 to 68 percent (for a reference, PRS ranks the Scandinavian countries at the top, while the median OECD has a percentile rank of 83 percent). On the World Bank Control of Corruption measure, Brazil went from a score of 57 out of 100 in 1996 to 54 in 2006, while Mexico increased from 36 to 47, and Chile rose slightly from 90 to 91 (for a reference, the median OECD country has a score of 93).

[2]This ranking includes all OECD countries plus Argentina, Brazil, Bulgaria, Costa Rica, Croatia, Cyprus, Indonesia, Kazakhstan, Malta, Romania, Russia, and South Africa.

[3]The choice of indicators is not obvious, as they show mixed results. For example, the Heritage Foundation property rights ranking shows an improving picture in Brazil, Mexico, and Chile.

Country-Specific Issues: Adoption, Timing, and Outcomes in Brazil, Chile, and Mexico

This section summarizes the adoption of Washington Consensus policies and the resulting outcomes in Brazil, Chile, and Mexico. Brazil only partially adopted the Washington Consensus reforms with unsatisfying results. Chile is certainly a poster child for early implementation and success, although a more nuanced view emerges in hindsight, especially of the recent past. Mexico implemented some of the Washington Consensus policies early on, but results have been rather disappointing. Table 1 illustrates our assessment of the degree of adoption of each of the ten principles of the Washington Consensus for the three case studies. In this section, we present further details on the adoption, timing, and results of the different policies in each of the three countries.

Brazil: Half-Hearted Adoption, Unsatisfying Outcomes

In the late 1980s, when the Washington Consensus debate appeared, Brazil was negotiating with creditors after defaulting on its debt. The economy was suffering from high inflation and bouts of hyperinflation. There was a widespread perception that the inward-oriented import substitution model—with substantial government intervention in the economy—had failed. The need to rein in inflation was the focus of policy efforts, which included a series of inflation stabilization plans: namely, the Cruzado plan of 1986, the Bresser plan of 1987, the Verão plan of 1989, and the Collor plan of 1990. These all failed to control high and hyperinflation, either because they lacked fiscal consolidation and monetary policy credibility or because they did not adequately deal with inflation inertia. Finally, the successful Real plan of 1994 solved these issues and led to a sequence of other reforms, several of which coincided with the Washington Consensus.

Brazil adopted the Washington Consensus reforms half-heartedly. Key early supporters of the Washington Consensus included influential former ministers and congressmen (including Mario Henrique Simonsen and Roberto Campos), who favored a smaller role for government, privatization of public companies, and less regulation. However, the perception that the Washington Consensus was a US idea and part of an IMF program conditionality led to a backlash. For example, Bresser-Pereira (1991) argued that it was necessary to overcome the fiscal crisis by reducing or canceling the public debt and recovering the savings capacity of the state. There were more balanced views, too. Malan (1991, p. 11, our translation) argued that "there is no single path, no simple formula or simple model to be followed. Each country in the region must analyze in-depth what it could be in the future. . .and adopt the 'appropriate policies.'"

Ultimately, Brazil partially adopted the Washington Consensus agenda, including fiscal consolidation (for a limited period of time), privatization, market-determined interest rates (despite substantial earmarked lending), and floating exchange rates (with exceptions, such as 1994–98).

Table 1
Washington Consensus Adoption in Brazil, Chile, and Mexico, 1990–2005

Williamson's Overarching Principles	*Brazil*	*Chile*	*Mexico*
1. Fiscal discipline (deficit of 1–2% of GDP)	○	●	◐
2. Public expenditure reallocation into priority sectors	◐	◐	◐
3. Broader tax bases and moderate marginal tax rates	◐	●	●
4. Market-determined and positive real interest rates	●	●	●
5. Competitive exchange rate, single regime	●	●	●
6. Trade liberalization, tariffs at 10–20% and low variance	○	●	●
7. Opening to foreign direct investment	●	●	●
8. Privatization to relieve public deficits and foster efficiency	◐	◐	◐
9. Deregulation to promote competition	◐	◐	◐
10. Property rights protection	◐	●	○

Source: Authors' assessment based on Figure 1, online Appendix (available at the *JEP* website), and text.
Note: White circles indicate low policy adoption and poor outcomes; gray, medium adoption and intermediate outcomes; and black, extensive adoption and strong outcomes.

The early steps toward fiscal consolidation included, in the late 1990s, a series of agreements with states and municipalities that capped the chronic spending and indebtedness of these local governments. Following a major fiscal adjustment in 1998–99, the approval of a Fiscal Responsibility Law in 2000 paved the way for 15 years of primary surpluses (several within IMF agreements), which helped stabilize public debt dynamics and the economy for some time. However, fiscal discipline has been gradually lost over the last decade or so, and local governments have created new rounds of budgetary troubles. Therefore, Brazil's legislative agenda continues to be dominated by the need for fiscal reforms, such as an overall spending cap (2017) and pension reform (2019). Other fiscal changes are currently under debate, including the administrative reform (on public sector wages and promotions).

The privatization process in Brazil continues to the present, but it has been slow and incomplete. The initial push was strong, with the telecommunications, banking, and mining sectors being privatized in the 1990s. The process continued with infrastructure concessions, the selling of oil field rights, and, more recently, the privatization of water and sewage companies. The government has also initiated efforts to privatize smaller companies, but it has not accepted selling the sacred cows, such as Petrobras (oil company) and Caixa and Banco do Brasil (banking sector).

There was substantial progress in financial liberalization, and the current perception is that interest rates and exchange rates are determined by the market. Several state banks were privatized in the mid-1990s. The government also liberalized the financial system and reduced public control of the banking sector. These conditions allowed interest rates to reach record low levels in 2017–20. Additionally, legislation in 2017 implemented market-oriented pricing in the national development bank (Banco Nacional de Desenvolvimento Econômico e Social or BNDES), allowing private capital markets to boom. Notwithstanding the financial

liberalization reforms, almost half of Brazil's credit is still government-directed lending (housing, agriculture, and BNDES), and two public banks are among the top five largest banks in the sector.

Trade openness, one of the main Washington Consensus reforms, was never adopted in Brazil. Unlike most of the rest of Latin America, Brazil remains one of the most closed economies in the world, due mostly to a political economic legacy of industries created under the import-substitution framework and the perception that there is a large domestic economy to defend. Despite some reduction in tariffs in the early 1990s during the short Collor government, tariffs and other barriers remain very high, and the only relevant trade agreement—Mercosur, with Argentina, Uruguay, Paraguay, and a few associates—has mostly diverted trade, rather than creating more of it.

The measures Brazil adopted in the 1990s were essential to stabilize rampant high inflation, avoid balance-of-payments crises, and prepare the economy to take advantage of the commodity boom of 2003–13, with gains in poverty alleviation. But the reforms were not enough to generate sustainable results. In particular, productivity growth performance has remained dismal.

Mexico: Early Implementation, Disappointing Results

Mexico suffered a severe economic and financial crisis in 1982. Authorities declared an external debt moratorium and nationalized the banks to stop the speculative attack against the peso. A new and more orthodox government took office shortly thereafter and embarked on an IMF-supported program, which included several aspects of the Washington Consensus, such as abandonment of the dual exchange rate regime, fiscal adjustment, some privatizations, and the beginning of the trade liberalization process, which included the incorporation of Mexico into the General Agreement on Tariffs and Trade (GATT) in 1985.

Despite the significant adjustment and the implementation of several Washington Consensus policies, another financial crisis took place at the end of 1987, with inflation peaking at 157 percent. This was probably the reflection of several events: a 1985 earthquake, the crash at the New York Stock Exchange in October 1987, and the impact of a significant drop in the oil price.

The Washington Consensus policies caused political controversy. The party that had ruled Mexico for 60 years, the *Partido Revolucionario Institucional* (PRI), ruptured in 1988, resulting in the split-off of the left-leaning *Partido de la Revolución Democrática* (PRD). One reason behind the split was a sense that the PRI had become less democratic, but the outward and market-oriented vision of the PRI also played a significant role (Márquez and Meyer 2010), especially because it meant that these policies were affecting the interests of very powerful groups.

A year later, with the beginning of a new government, the country launched an ambitious reform program as part of a stabilization plan that included an agreement on the trend of price adjustments among labor unions, the private sector, and the government. The program incorporated important elements of the Washington Consensus. In less than five years, authorities privatized leading state-owned

companies; enacted a fiscal reform aimed at increasing the tax base and reducing marginal tax rates, while significantly reducing the fiscal deficit; liberalized the financial system and the financial account (both foreign direct investment and portfolio); and reduced barriers to entry in strategic sectors. In some cases, these policies went beyond the Washington Consensus, while in others the recommendations were implemented only partially. One important reform was trade liberalization during the first half of the 1990s, which culminated with the signing of the North American Free Trade Agreement as well as other free trade treaties.

The opening to international competition contributed to macroeconomic stabilization as well as to market discipline in the tradable sector. Moreover, the manufacturing sector registered double-digit annual growth rates for more than a decade and significant increases in productivity. However, the nontradable sector in Mexico has been less dynamic. Despite some flexibility, private investment in the energy sector remained very restricted until 2013. This translated into low investment rates, declining productivity, and even lower production. In other services, such as telecommunications and transport, there is still significant room for improvement through deregulation, the implementation of adequate regulation to facilitate business operation, the reduction of barriers to entry, and the fight against monopoly power.

In the last two decades, Mexico's growth has been disappointing, productivity has increased very slowly, and real wages have remained almost flat. One extreme view is that this lack of progress is due to the Washington Consensus model itself; at the opposite extreme, others argue that the reforms were not deep enough (Gil Díaz 2003) or that the implementation was weak (Cordera and Lomelí 2002). Another argument is that the Washington Consensus left out relevant issues (Grupo Huatusco 2004). External shocks, especially the expansion of China in world trade and its impact on manufacturing and commodity prices, had a negative effect on Mexico's terms of trade, while other Latin American economies benefited as commodity exporters. According to Levy (2018), one significant limitation to economic growth is the perverse incentives that persist in the labor market. The relatively high taxes and social security contributions in the formal sector generate a large and increasing informal sector characterized by low productivity and wages. Additionally, the persistent low quality of education in Mexico, even as it has improved in other emerging markets, has severely limited the accumulation of human capital. For a long time, the teachers' labor union was powerful enough to stop any attempts to reform. It was not until 2013 that the government took a step in the right direction, but a counter-reform in 2019 eliminated the fundamental changes. Broad access to quality public education remains pending. Public expenditure and investment are still very inefficient (Esquivel 2003; Izquierdo et al. 2017).

Finally, a critical factor has been Mexico's dreadful performance on property rights or, more generally, the enforcement of the rule of law—perhaps the weakest aspect the of country's economic institutions. Even by Latin American standards, Mexico stands out for the level of corruption, the lack of access to justice for most of the population, the rampant power of mafias, and the weak protection of property

rights, among other issues related to the weak legal institutions. There has been progress with the publication of a new bankruptcy law in 2000, the constitutional reform of the judicial system of 2008, which deeply transformed the Mexican legal system and the constitution of specialized courts for antitrust and telecommunications cases. These efforts have clearly been insufficient as most of the legal and judicial system indicators have worsened in recent years. There is overwhelming agreement regarding the negative impact of the weak legal framework not only on economic growth, but also on the quality of life of the Mexican people.

Chile: Success, but Less So in Hindsight

There was very little political opposition to the Washington Consensus in Chile, partly because it was implemented under a military dictatorship. Some elements of the Washington Consensus, like greater security for property rights, trade integration, privatization, and openness for direct investment, had been implemented in the 1970s. For example, of the 570 companies that the state controlled in 1973, only 24 were still publicly held in 1983. After a brief stint with heterodox policies after a deep economic crisis in 1982–83, Chile adopted almost all the Washington Consensus policies. Meller (1990, 1996) reports that the Chilean economic team that took control in 1985 was considered a more avid fan of the IMF than even the IMF itself. Privatizations in 1985–88, the tax reform of 1986, and policies to support a competitive exchange rate were fundamental.

After Chile's transition to democracy in 1989, the first (center-left) democratic government continued to embrace the Washington Consensus. Trade integration, increasing exchange rate flexibility, and prioritizing spending on social needs became landmarks of economic policy. Moreover, John Williamson was seen as somewhat progressive (and a friend) by local economists, so the Washington Consensus was not perceived as a US imposition. Productivity increased vigorously in 1987–2010, notably in the first decade, led by foreign direct investment in mining and the development of new export sectors. Since the mid-1990s, macroeconomic policies have remained well-aligned with best practices, including the adoption of a full-fledged inflation-targeting regime, a floating exchange rate, and a fiscal rule. Macroeconomic stability is now basically taken for granted. The country also made progress with infrastructure investment through public–private partnerships and new social strategies, such as a public system of health guarantees, unemployment insurance, a minimum pension scheme, and many education reforms. A few economists criticized the floating exchange rate regime and financial integration (for example, Ffrench-Davis 2005), and there has recently been some political push-back against privatized public services—especially toll roads, which are considered expensive, and any public service that suffers an interruption—but there have been no serious attempts to reverse any of these policies.

Changing spending priorities and deregulation were the only two elements not fully implemented during the Pinochet military dictatorship and later on during democracy, although there was some progress. Spending was duly concentrated on social needs after 1990, but it remained limited relative to the size of the

economy. Partly due to the small size of the domestic market and a history of large economic family-owned conglomerates, ownership continued to be quite concentrated. Privatized companies also ended up in only a few hands, and though foreign direct investment expanded significantly, it has been concentrated in mining and nontradable industries where it is more difficult to have several players. Regulations fostered competition in some areas (like telecoms), but they were not as effective in others (like fisheries and the private pensions system). Developing a strong and independent antitrust agency took almost two decades.

Despite evident economic progress over the last three decades, Chile suffered severe social unrest in October 2019. In response to the widespread protests and violence, the main political parties agreed to a referendum vote on a new constitution to be written by an elected assembly in 2021. There are competing theories as to why so many Chilean citizens became fed up with the government, politicians, and institutions. One hypothesis is that relatively low per capita growth in the last few years, coupled with substantial immigration, stressed a large but still vulnerable middle class. Another explanation is that social tensions gradually accumulated as citizens' priorities changed while the social contract was overly slow to adapt. UNDP (2017) summarizes the findings of their annual reports of the last 20 years as follows: "in 1999, Chileans mainly dreamed of becoming an economically developed country; in 2016, they dreamed of having a safer, more protective, and fairer country" (p. 32).

Chile has built an excessively unequal society behind its apparent macroeconomic success. Strong growth helped poverty decline very quickly, while an emerging middle class expanded. However, besides a poor and slowly improving income distribution, there are limited risk-sharing arrangements and a widespread perception of unfair procedures given the country's income level. For example, the core of the pension system is based on individual capitalization accounts, and there is a two-tier health system, with a state-managed, low-quality tier for 80 percent of the population and a more developed tier for the wealthiest 20 percent. In contrast to many developed countries, Chilean cities and education are quite segregated. The middle class feels overindebted after having massive access to credit. There is low penetration into a wealthy and powerful elite (Zimmerman 2019), and there is a perception of vast impunity for the elite's wrongdoings. Additionally, markets appear to be too concentrated, competition in specific industries is weak, and some businesses have proved to be too intertwined with politics. Some of these shortcomings are unrelated to the core of the Washington Consensus, but some do relate to better regulation, more competition, and public spending volume and priorities.

Performance after Three Decades: Improvements in Inflation and Poverty but Dismal Productivity Growth

Our description of the evolution of the Washington Consensus in Brazil, Mexico, and Chile illustrates some of the difficulties in evaluating the "consensus."

It is challenging to untangle the effects of several other policy initiatives, the lack of proper implementation, and external shocks. Commodity cycles, for example, have definitely been quite relevant for the region's short-run growth performance in certain periods. These problems are magnified three decades later: improvements in some of the indicators in the latter part of the period probably do not reflect the direct impact of the Washington Consensus policies, but rather derive from new policy agendas. One could argue, however, that the Washington Consensus policies may have set the stage for the new agenda and thus had an indirect impact on the outcomes.

In this section, we discuss the economic performance of Latin America along various dimensions since the 1980s. Outcomes from the 1990s, in particular, tend to have a more direct connection to the Washington Consensus policies, while outcomes since about 2000 are progressively influenced by additional policies and events. Overall, Latin America made progress in reducing inflation and, since 2000, poverty, but growth/productivity performance was generally poor. Table 2 summarizes our assessment of the key outcomes in our three countries for the full period.

Inflation

One important achievement of the Washington Consensus policies was taming inflation. The median annual inflation rate in Latin America was 100 percent in the 1980s, with occasional hyperinflation well above that level. The median inflation rate fell to about 40 percent in the 1990s, and it has been 5–6 percent per year since 2000 (based on IMF data). Inflation volatility also declined significantly in the 1990s—progress that remains today. Very few Latin American countries still regard high inflation as a primary concern.

Several countries have consolidated these gains against inflation by legislating or granting functional central bank independence and also adopting successful inflation-targeting regimes. This went beyond the original Washington Consensus recommendations, following newer best practices in monetary policy. For example, according to the Garriga (2016) index of central bank independence, Chile, Colombia, Mexico, and Peru increased their central bank independence significantly in the 1990s. Central bank independence and inflation targeting gained importance as Latin American countries moved toward a more flexible exchange rate regime. By the end of the 1990s, this became a cornerstone of greater macroeconomic stability in many Latin American countries.

Growth and Productivity

Latin America's growth performance in the last three decades improved relative to the 1980s, but it has fallen short of expectations at the outset of the Washington Consensus and has been consistently poor relative to other emerging markets. Regional real per capita GDP (measured using purchasing power parity exchange rates) *declined* –0.4 percent per year during the lost decade of the 1980s, and it has grown 1.2 percent per year since 1990. For comparison, per capita GDP in advanced economies gained 1.3 percent in the 1980s and 1.5 percent since 1990, while in emerging markets as a group, per capita growth accelerated from

Table 2
Performance in Brazil, Chile, and Mexico, 1990–2020

Key outcomes	*Brazil*	*Chile*	*Mexico*
Productivity growth	○	●	○
Inflation	●	●	●
Change in poverty	●	●	◉
Change in income distribution	◉	◉	◉

Note: White circles indicate a poor outcome; gray, intermediate; and black, strong.

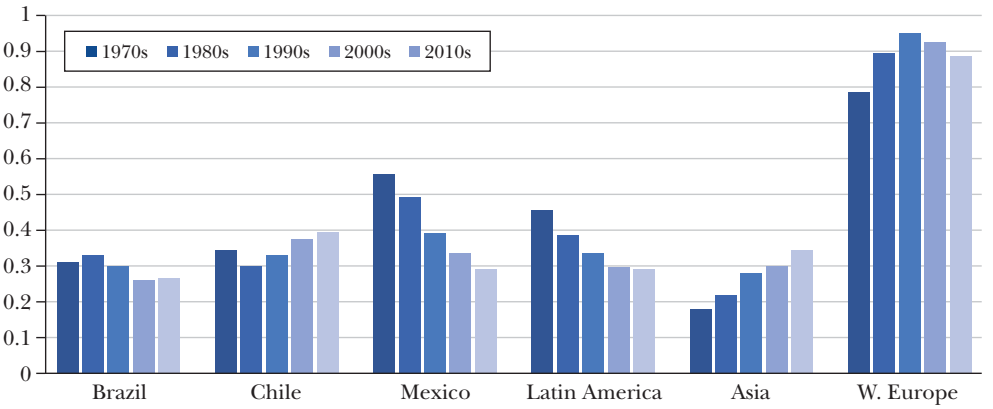
1.2 percent annually in the 1980s to 3.2 percent since 1990. In Latin America, only Chile had a higher growth rate than the average for emerging markets.

Overall, the evidence suggests that countries that more fully adopted the Washington Consensus policies generally had a better growth performance. For example, in this journal, Fraga's (2004) early evaluation of the Washington Consensus finds that the Latin American countries that were more active in carrying out the consensus reforms also experienced better economic performance, whereas Ocampo (2004) offers a nuanced view, worrying particularly about procyclical macroeconomic policies and weak productivity growth. In more recent studies, Estevadeordal and Taylor (2013) find a positive and significant impact of trade liberalization on economic growth. Easterly (2019) presents three stylized facts that cast doubts on the alleged failure of the Washington Consensus policies to foster growth. Grier and Grier (2020) show that Washington Consensus policies did reliably raise average incomes: countries that had sustained reform were 16 percent richer ten years later. In our case studies, Chile performed well, while more mixed adopters, such as Brazil and Mexico, underperformed.

Of course, long-run growth is necessarily built on productivity. Latin America has had an endemic shortfall of savings and investment, a situation that did not change with the Washington Consensus. From 1980 to 2019, emerging market economies worldwide averaged a savings and investment rate of 27 percent of GDP (according to IMF data). Over the last four decades, Brazil and Mexico remained significantly below the emerging market average on both fronts. Chile had a few periods with higher investment, especially in the 1990s, but these bouts were short-lived. However, Bakker et al. (2020) conclude that total factor productivity, rather than investment ratios, explains the slow income convergence of Latin America and the Caribbean in comparison with Emerging Europe.

Productivity requires human capital accumulation. According to OECD data, Brazil, Chile, and Mexico increased expenditure on education between 1 and 2 percentage points of GDP between 1990 and the early 2000s. This trend continued in the following decade. Though available data are more sporadic for other countries, the overall picture for Latin America is similar. As a result, Latin America's gross enrolment rate in secondary education increased from 77 percent in 1990 to 85 percent in 2000 and 89 percent in 2010. In our three countries, secondary

Figure 2
Productivity: Output per Hour Worked Relative to the United States



Source: The Conference Board Total Economy Database
Note: Latin America: Simple average of Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Venezuela. Asia: Simple average of China, Hong Kong, India, Indonesia, Japan, Malaysia, Myanmar, Pakistan, Philippines, Singapore, South Korea, Sri Lanka, Taiwan, Thailand, and Vietnam. Western Europe: Simple average of Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and United Kingdom.

education for the 20- to 24-year old population increased to an average of 3.7 years by 2010, 1.4 years more than in 1990. But the quality of education remains poor. For example, in the 2000 Programme for International Student Assessment (PISA) test, Latin American countries were at the bottom of the results, even below what should be expected given per capita income. In the 2009 PISA round, the region improved in reading but was still low in the rankings.

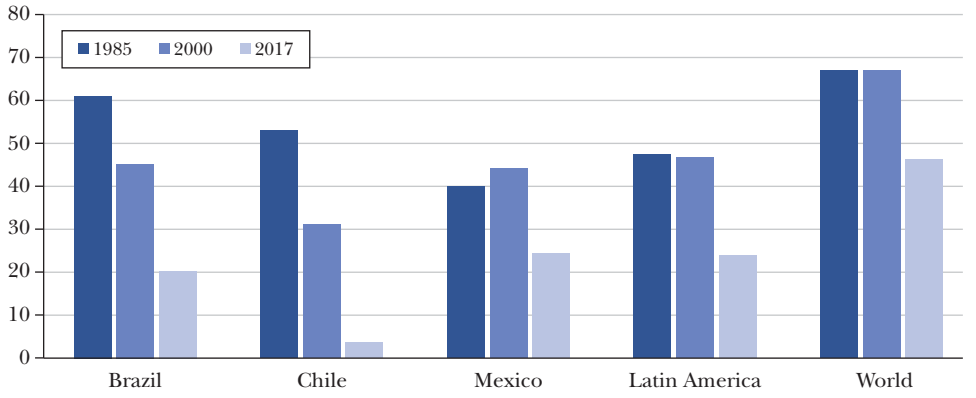
Not surprisingly, the growth of output per hour in Latin America relative to the United States declined, in marked contrast to Asia, as shown in Figure 2. In the region, only Chile managed some relevant catch-up in the last 30 years. Mexico is perhaps the most puzzling: it continued its previous relative declining trend after the Washington Consensus. Brazil also had a poor performance.

Interestingly, the countries that are perceived to have closely followed Washington Consensus policies—namely, Chile, Colombia, and Peru—had a better performance in the last three decades in terms of reversing the decline of the 1970s and 1980s to a degree. The countries that departed the most from the Washington Consensus, like Argentina and Venezuela, recorded a poor growth performance, as well as high volatility.

Poverty

Reducing poverty rates was not an explicit goal of the Washington Consensus policies. Based on the World Bank poverty line of US$5.50 per day for upper-middle-income countries, the Latin American region in general had essentially no decline

Figure 3
Poverty Head Count
(percent of total population)



Source: World Bank World Development Indicators.
Note: Poverty is defined as US$5.50 per person per day, at 2011 purchasing power parity exchange rates.

in poverty rates from 1985 to 2000. Figure 3 shows that poverty rates fell substantially from 2000 to 2017 in the region, reflecting the benefits of the commodity boom, higher growth in some countries, and targeted transfer programs in others.

In the late 1990s, government expenditure in Latin America was reallocated to social programs to reduce poverty and increase social mobility. Countries moved away from general food subsidies and guaranteed prices for essential crops, shifting to conditional cash transfer programs that target the most disadvantaged segments of the population, an instrument which was not part of the Washington Consensus. Brazil and Mexico both developed this type of national poverty alleviation program. The names of the programs have changed with new governments—Bolsa Escola/Bolsa Familia in Brazil and Progresa/Oportunidades/Prospera in Mexico—but the programs themselves remain firmly in place. Several other Latin American countries, including Argentina, Colombia, Ecuador, Honduras, and Nicaragua, developed similar cash transfer programs. Formal program evaluations suggest a significant increase in school attendance (Rawlings and Rubio 2005).

Despite the benefits of most of these programs on poverty, the intergenerational transmission of poverty has been only marginally reduced in the last 20 years or so. Programs have focused mainly on solving access problems, without any direct effect on supply or quality shortcomings. Thus, deficiencies in the quality of education, health services, and even iron supplements have affected the long-term impact of the programs (Lomelí 2008).

Income Distribution

Although income distribution remains unequal and is an important issue going forward, there was more progress than what the limited GDP and productivity growth would suggest. Indeed, income distribution improved in the region relative to the trend in many industrialized countries, though not necessarily as a result of the Washington Consensus policies. Across Latin America, income growth of the bottom 20 percent and the middle class was significantly higher than that of the wealthiest 20 percent, both in the 1990s and afterward (Table 3). In Brazil, Chile, and Mexico, redistribution was as important as growth for the poorest 20 percent.

Income was not the only area of welfare progress. Working hours have declined in the region, in line with the standard relation between hours and income. In Mexico, working hours remain somewhat above the norm, whereas in Brazil, they are below. Life expectancy in the region increased on par with the world, with Mexico lagging in the last decade.

The Washington Consensus in Latin America and its Aftermath

Since the inception of the Washington Consensus in the late 1980s and early 1990s, Latin American economies have become significantly more stable, with less frequent instances of balance-of-payments crises, high or hyperinflation, and unsustainable debt dynamics. However, it is fair to conclude that Latin American economic performance has been disappointing over the last 30 years, both compared with other regions and emerging economies and relative to expectations at the beginning of the 1990s. Even success cases, like Chile, are currently under scrutiny. Growth performance improved relative to the lost decade of the 1980s, but forecasts and targets made over the years, including by the IMF, show substantial economic underperformance.

How much of this outcome can be attributed to—or occurred despite—the Washington Consensus reforms is still under debate. There is also controversy as to whether the Washington Consensus principles were actually implemented. No economy took all the recommendations fully, and most of the countries were either slow or not persistent in adopting them. But a substantial share of the countries in Latin America did adopt at least a reasonable subset of the initial recommendations. Over the years, many countries increasingly took on board fiscal responsibility (albeit imperfectly), inflation control, floating exchange rates, market-determined interest rates, privatization, trade openness, and spending on education and health.

Certainly, some important aspects of Washington Consensus policies have been successfully implemented and provide important building blocks for a successful development strategy and model. For example, trade openness became widely shared in countries like Chile, Colombia, Mexico, and Peru. In Mexico in 1993, there were intense demonstrations against the North America Free Trade Agreement, and left-wing parties opposed it in the Mexican Senate. In contrast, the Mexican Senate's approval of the new United States-Mexico-Canada Agreement

Table 3

Per Capita Income by Income Share in 1990, 2004, and 2018

(at purchasing power parity exchange rates, in 2018 US dollars and percent)

Country or region and income segment	Per capita income (%)			Annual growth rate (%)		Due to redistribution (%)	
	1990	2003	2018	1990–2003	2003–2018	1990–2003	2003–2018
Brazil							
Average	12,071	13,116	16,146	0.6	1.4	—	—
Poorest 20%	1,388	1,705	2,503	1.6	2.6	0.9	1.2
Middle 60%	6,659	7,782	10,387	1.2	1.9	0.6	0.5
Richest 20%	38,991	40,528	47,066	0.3	1.0	–0.3	–0.4
Chile							
Average	10,163	17,182	25,700	4.1	2.7	—	—
Poorest 20%	1,728	3,436	7,453	5.4	5.3	1.3	2.5
Middle 60%	5,776	11,226	18,376	5.2	3.3	1.1	0.6
Richest 20%	31,759	48,798	65,922	3.4	2.0	–0.7	–0.7
Mexico							
Average	14,620	17,314	20,616	1.3	1.2	—	—
Poorest 20%	2,632	3,636	5,566	2.5	2.9	1.2	1.7
Middle 60%	9,064	11,629	14,741	1.9	1.6	0.6	0.4
Richest 20%	43,275	48,047	53,293	0.8	0.7	–0.5	–0.5
Average of:							
Latin America	11,280	12,725	16,224	0.9	1.6	—	—
Emerging and developing	5,137	6,701	12,510	2.1	4.2	—	—
Advanced economies	34,326	43,124	51,776	1.8	1.2	—	—

Source: Authors' estimations based on International Monetary Fund World Economy Database and World Bank World Development Indicators.
Note: We calculate the per capita GDP for each subgroup with income distribution data and the redistribution effect by subtracting the (compounded) average per capita GDP growth.

(USMCA) in 2019 was almost unanimous, with barely any demonstrations. In Chile, negative sentiment to the Trans-Pacific Partnership relates more to the polarization of anti-American (and pro-China) groups than to a push for closing the economy. Brazil is the main exception to this movement toward greater trade openness, and its economy remains closed.

In addition, despite the current fiscal challenges in Brazil and other countries, there is general agreement in Latin America on the importance of maintaining fiscal discipline and keeping control of government indebtedness to avoid macroeconomic instability. Even in Mexico, now governed for the first time by a left-wing party, President López Obrador's commitment has been strong: "We are going to maintain no fiscal deficit, no matter what."[4] Across Latin America there are now

[4] "Vamos a mantenernos sin déficit fiscal pase lo que pase, asegura AMLO en convención bancaria," El CEO, March 13, 2020 (our translation). See https://elceo.com/economia/vamos-a-mantenernos-sin-deficit-fiscal-amlo-convencion-bancaria/.

fiscal policy rules, fiscal responsibility laws, independent fiscal councils, and explicit mechanisms for evaluating policies both before and after they are enacted. A number of regulatory and supervisory institutions have also been strengthened over the last three decades. One of the factors driving institutional modernization is the region's increasing integration in multilateral organizations: the OECD incorporated Mexico as a full member in 1994, Chile in 2010, and recently Colombia, while Brazil is the most active OECD "key partner" (as a nonmember). Current fiscal policy discussions in the region are centered on the right level of debt and deficits and the need for countercyclical fiscal policy, though COVID-19 will leave countries with severe budgetary challenges.

In other areas, the legacy of the Washington Consensus is being questioned. Privatization and the role of the state remains a divisive issue in some countries. The speed of privatizations in the 1990s came at the sacrifice of putting in place an adequate regulatory and supervisory scheme to allow competition in the newly privatized sectors. In Mexico, the current government has blocked private investment in the energy sector. In Chile, new privatizations are out of the question. There is also debate about deregulation, as some consider the government's regulatory capacity to be limited and fear new monopolistic powers.

In addition, the Washington Consensus policies were delineated during a time of debt crisis and severe macroeconomic stress and thus fell short of a full development strategy. To be sure, countries were able to move ahead on other policy agendas, including strengthening institutions (for example, central banks), pension and savings reforms, and social policies, such as conditional transfers for the most vulnerable. However, despite the reduction in poverty and some improvement in income distribution, the advances in social areas in the last decades are considered insufficient, and there is a perception that more is urgently needed. Targeted government spending in education and health has been more noticeable since the 2000s, but there is still a long way to go in terms of quality and fair access.

Several important areas of public concern, which were not part of the Washington Consensus (and not even considered major issues at the time), are becoming critical: i) public security and the fight against organized crime, usually related to drugs (Latin America and the Caribbean represents 8 percent of the world population, but has more than 40 percent of world homicides); ii) access to justice, as citizens feel that elites receive preferential treatment; iii) corruption, which has deteriorated significantly in recent years and has had a heavy toll on the credibility and legitimacy of politicians; and iv) environmental policies, particularly in Brazil, with the debate on conservation of the Amazon rain forest, and more recently in Mexico, with the debate on green versus traditional energy.

The Washington Consensus seems likely to remain a subject of controversy. On one side, it bears the burden of a number of negative assessments (for example, Rodrik 2006). Stiglitz (2008, p. 41) provides a summary of the critical view: "There is no consensus except that the Washington Consensus did not provide the answer." On the other side, Grier and Grier (2020) argue that the alternatives to the Washington

Consensus have performed even worse. Easterly (2019) concludes that the evidence "seems most consistent with a position in between the poles of complete dismissal or vindication of the Washington Consensus" (p. 35).

In current public policy debates in Latin America, controversy over "neoliberalism" dwarfs interest in the Washington Consensus. Neoliberalism is the straw man most commonly held up as responsible for Latin America's economic problems. According to our calculations using the Google Books Ngram Viewer, books published in 2019 in Spanish had 70 times more references to "neoliberalism" than to the "Washington Consensus."

But neoliberalism is not a clearly defined concept in economics. In public discussion, neoliberalism is narrowly associated with a laissez-faire view (à la Hayek) and perhaps also with extreme monetarism (à la Friedman), and it is sometimes equated with rather orthodox and pro-market reforms. Neoliberalism has also been identified with policies that disregard some relevant aspects of development, such as inequality and poverty, and neglect any role for the state. More importantly for the issues discussed here, critics have sometimes caricatured the Washington Consensus as a neoliberal manifesto. As described by Thorsen (2010, p. 3), neoliberalism has become "a generic term of deprecation to describe almost any economic and political development deemed undesirable." The Washington Consensus should not be mechanically associated with this neoliberal straw man. As shown in this paper, the Washington Consensus was a list of recommendations that was partially adopted with mixed results, some of which were satisfactory and others clearly not.

In our view, without some subset of the Washington Consensus policies, it would have been difficult, if not impossible, to achieve macroeconomic stability and to recover access to foreign financing in the late 1980s and early 1990s. The main risk in Latin America at present is that economic populism will gain ground and policymakers will discard the Washington Consensus policies altogether. That would be a mistake. The reality is that many of the Washington Consensus policies are needed as building blocks for a new agenda. Whatever the merits are of the Washington Consensus policy agenda in the last three decades, Latin America in the 2020s faces a larger set of policy challenges, including social, income distribution, education, security, rule of law, and environmental issues.

■ *We thank the JEP Editors; Guillermo Babatz, José De Gregorio, Arminio Fraga, Pablo García, Elena Landau, Aaron Tornell, Patricio Meller, and Alejandro Werner for useful comments; the Department of Statistics at Banco Central do Brasil; and João Luiz Queiroz da Silva Ayres, Solange Srour, and Lucas Vilela for help with information and data.*

References

Aspe, Pedro. 1993. *Economic Transformation the Mexican Way*. Cambridge, MA: MIT Press.

Baker III, James A. 1985. "Statements by Governors: James A. Baker III." In *1985 Annual Meetings of the Annual Meetings of the Boards of Governors Summary Proceedings*, 205–14. Washington, DC: International Bank for Reconstruction and Development, International Finance Corporation, and International Development Association.

Bakker, Bas B., Manuk Ghazanchyan, Alex Ho, and Vibha Nanda. 2020. "The Lack of Convergence of Latin America Compared with CESEE: Is Low Investment to Blame?" IMF Working Paper 20/98.

Banco Central do Brasil, Department of Statistics. 2021. "Resultados Nominal e Operacional Séries. xlsx". Available upon request at dstat@bcb.gov.br.

Barro, Robert J. 2015. "Convergence and Modernization." *Economic Journal* 125 (585): 911–42.

Borensztein, Eduardo, Kevin Cowan, Ugo Panizza, and Barry Eichengreen. 2008. *Bond Markets in Latin America: On the Verge of a Big Bang?* Cambridge, MA: MIT Press.

Bresser-Pereira, Luiz Carlos. 1991. "A crise da América Latina: Consenso de Washington ou crise fiscal?" *Pesquisa e Planejamento Econômico* 21 (1): 3–23.

Chong, Alberto, and Florencio López-de-Silanes. 2005. "The Truth about Privatization in Latin America." In *Privatization in Latin America: Myths and Reality*, edited by Alberto Chong and Florencio López-de-Silanes, 1–66. Palo Alto, CA: Stanford University Press.

Cordera, Rolando, and Leonardo Lomelí. 2002. "Viejos y nuevos paradigmas: El papel político de las ideas económicas en el cambio estructural en México (1982–1994)." Unpublished.

Díaz, Ela, and Rodrigo O. Valdés. 2020. "All that Glitters Is Not Gold: A Ranking of Global Rankings." *Economía* 20 (2): 223–54.

Easterly, William. 2019. "In Search of Reforms for Growth: New Stylized Facts on Policy and Growth Outcomes." NBER Working Paper 26318.

Esquivel, Gerardo. 2003. "México: En pos del crecimiento." Centro de Estudios Económicos Working Paper VIII.

Estevadeordal, Antoni, and Alan M. Taylor. 2013. "Is the Washington Consensus Dead? Growth, Openness, and the Great Liberalization, 1970s–2000s." *Review of Economics and Statistics* 95 (5): 1669–90.

Ffrench-Davis, Ricardo. 2005. "Reforming the Reforms: Why and How." In *Reforming Latin America's Economies after Market Fundamentalism*, 1–24. London: Palgrave Macmillan.

Fraga, Arminio. 2004. "Latin America since the 1990s: Rising from the Sickbed?" *Journal of Economic Perspectives* 18 (2): 89–106.

Garriga, Ana Carolina. 2016. "Central Bank Independence in the World: A New Dataset." *International Interactions* 42 (5): 849–68.

Gil Díaz, Francisco. 2003. "Don't Blame Our Failures on Reforms That Have Not Taken Place." *Fraser Forum*. June 2003, 7–11.

Grier, Kevin B., and Robin M. Grier. 2020. "The Washington Consensus Works: Causal Effects of Reform, 1970–2015." *Journal of Comparative Economics* 49 (1): 59–72.

Grupo Huatusco. 2004. *¿Por qué no crecemos? Hacia un consenso para el crecimiento en México.*

Ilzetzki, Ethan, Carmen M. Reinhart, and Kenneth S. Rogoff. 2019. "Exchange Arrangements Entering the Twenty-First Century: Which Anchor Will Hold?" *Quarterly Journal of Economics* 134 (2): 599–646.

International Monetary Fund. 2020. "World Economic Outlook Database April 2020." https://www.imf.org/en/Publications/WEO/weo-database/2020/April (accessed March 1, 2021).

Izquierdo, Alejandro, Ruy Lama, Juan Pablo Puig, Daniel Riera-Crichton, Carlos A. Végh, and Guillermo Vuletin. 2017. "On the Determinants of Public Investment Multipliers." Unpublished.

Levy, Santiago. 2018. *Under-Rewarded Efforts: The Elusive Quest for Prosperity in Mexico*. Washington, DC: Inter-American Development Bank.

Levy, Yeyati, Eduardo, and Federico Sturzenegger. 2016. "Classifying Exchange Rate Regimes: 15 Years Later." Center for International Development Working Paper 319.

Lomelí, Enrique Valencia. 2008. "Conditional Cash Transfers as Social Policy in Latin America: An Assessment of Their Contributions and Limitations." *Annual Review of Sociology* 34: 475–98.

Lora, Eduardo. 2001. "Structural Reforms in Latin America: What Has Been Reformed and How to Measure it." Inter-American Development Bank Working Paper 466.

Malan, S. Pedro. 1991. "Uma Crítica ao Consenso de Washington." *Revista de Economia Política* 11 (3): 5–12.

Márquez, Graciela, and Lorenzo Meyer. 2010. "Del autoritarismo agotado a la democracia frágil, 1985–2009." In *Historia general de México ilustrada*, vol. 2, edited by Erik Velásquez García, Enrique Nalda, Pablo Escalante Gonzalbo, Bernardo García Martínez, Bernd Hausberger, Óscar Mazín, Dorothy Tanck de Estrada et al., 448–511. Mexico City: Colegio de México, Cámara de Diputados.

Mauro, Paolo, Rafael Romeu, Ariel Binder, and Asad Zaman. 2013. "A Modern History of Fiscal Prudence and Profligacy." IMF Working Paper 13/5.

Meller, Patricio. 1990. "Revisión del proceso de ajuste chileno de la década del 80." *Colección Estudios Cieplan* 30: 5–54.

———. 1996. *Un siglo de economia política chilena (1890–1990)*. Santiago: Andrés Bello.

Ocampo, José Antonio. 2004. "Latin America's Growth and Equity Frustrations during Structural Reforms." *Journal of Economic Perspectives* 18 (2): 67–88.

OECD Tax Database 2021. OECD. https://www.oecd.org/tax/tax-policy/tax-database/.

Rawlings, Laura B., and Gloria M. Rubio. 2005. "Evaluating the Impact of Conditional Cash Transfer Programs." *World Bank Research Observer* 20 (1): 29–55.

Rodrik, Dani. 2006. "Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's 'Economic Growth in the 1990s: Learning from a Decade of Reform.'" *Journal of Economic Literature* 44 (4): 973–87.

Stiglitz, Joseph E. 2008. "Is There a Post Washington Consensus Consensus?" In *The Washington Consensus Reconsidered: Towards a New Global Governance,* edited by Narcís Serra and Joseph E. Stiglitz, 41–56. Oxford: Oxford University Press.

The Conference Board. 2019. "Total Economy Database™ April 2019." https://conference-board.org/data/economydatabase/total-economy-database-archive (accessed July 20, 2019).

Thorsen, Dag Einar. 2010. "The Neoliberal Challenge: What Is Neoliberalism?" *Contemporary Readings in Law and Social Justice* 2 (2).

Tornell, Aaron, Frank Westermann, and Lorenza Martínez. 2003. "Liberalization, Growth, and Financial Crises: Lessons from Mexico and the Developing World." *Brookings Papers on Economic Activity* 2003 (2): 1–112.

Trading Economics. 2021. Trading Economics. https://tradingeconomics.com/country-list/corporate-tax-rate.

Williamson, John. 1990a. *Latin American Adjustment: How Much Has Happened?* Washington, DC: Peterson Institute for International Economics.

———. 1990b. "What Washington Means by Policy Reform." In *Latin American Adjustment: How Much Has Happened?*, edited by John Williamson. Washington, DC: Peterson Institute for International Economics.

World Bank. 2020. "World Development Indicators." https://databank.worldbank.org/source/world-development-indicators (accessed March 10, 2021).

UNDP (United Nations Development Program). *2017. Chile en 20 años: Un recorrido a través de los Informes sobre Desarrollo Humano*. Santiago: UNDP.

Zimmerman, Seth D. 2019. "Elite Colleges and Upward Mobility to Top Jobs and Top Incomes." *American Economic Review* 109 (1): 1–47.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 133–156

Washington Consensus Reforms and Lessons for Economic Performance in Sub-Saharan Africa

Belinda Archibong, Brahima Coulibaly, and Ngozi Okonjo-Iweala

The story of how African countries experienced a debt crisis in the 1980s began in the 1960s and 1970s, when newly independent African governments, struggling to help their countries recover from the ravages of European colonialism, carried out expansionary fiscal spending aimed at economic development. Governments also borrowed significantly to finance development expenditures over this time. Particularly for African countries that were oil exporters, the high oil prices of the 1970s made this borrowing look affordable. But in the 1980s, the economic tides turned. Falling oil prices along with a collapse in world prices of primary agricultural commodities, which made up 88 percent of Africa's exports, resulted in a shortfall in export revenues that put enormous pressure on government finances (Onyekwena and Ekeruche 2019; Mkandawire and Soludo 1999). The early 1980s also saw a global recession, along with an increase in real interest rates in donor countries that raised interest payments on previously contracted US dollar-denominated loans, markedly increasing the debt burden of African countries (Onyekwena and Ekeruche 2019; Due and Gladwin 1991). Additionally, Africa's governments featured largely in domestic financial institutions like

■ *Belinda Archibong is Assistant Professor of Economics at Barnard College, Columbia University, New York City, New York. Brahima Coulibaly is Vice President of the Global Economy and Development, Brookings Institution, Washington, DC. Ngozi Okonjo-Iweala is Director-General of the World Trade Organization, Geneva, Switzerland. She is also a former Finance Minister of Nigeria; Nonresident Distinguished Fellow with the Global Economy and Development Program, Brookings Institution, Washington, DC; and Chair of the Board of the Global Alliance for Vaccines and Immunization, Washington, DC.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.133.

the banking sector, with many of them nationalizing foreign banks or creating new state-owned financial institutions (Mkandawire 1999).

By the early and mid-1980s, many African governments were in severe financial straits and with lowered incomes, increasing poverty, and declining welfare, they turned to international financial institutions for debt relief. When economist John Williamson (1993) coined the term "Washington Consensus" in 1989, he was referring to a set of ten market-oriented policies that were popular among Washington-based policy institutions at that time, particularly as prescriptions for improving economic performance in Latin American countries. These policies centered around fiscal discipline, market-oriented domestic reforms, and openness to trade and investment. For indebted African countries, the Washington Consensus inspired market-based "structural adjustment programs" prescribed by international financial institutions, like the World Bank and the International Monetary Fund (IMF), that were often prerequisites for financial assistance (Onyekwena and Ekeruche 2019; Naiman and Watkins 1999; Mkandawire and Soludo 1999). Several African countries adopted these market-oriented policies beginning in the 1980s. The number of reform adopters increased further following the introduction of the Highly Indebted Poor Countries (HIPC) initiative by the World Bank and IMF in the mid-1990s, which provided debt relief to countries with "unsustainable" debt, provided they enacted many of the structural adjustment policies (Onyekwena and Ekeruche 2019).

It has been over three decades since these policies were first adopted across Africa and other developing countries, yet the evidence of their impact on economic outcomes remains a subject of debate. In this essay, we begin with an overview of the earlier evidence on the effects of these policies, which sometimes emphasizes the importance of policy inputs that go beyond the reforms themselves, like government capacity and public support. We then revisit whether market-oriented reforms of the 1980s and 1990s may have contributed to later positive economic outcomes for sub-Saharan Africa, with a focus on descriptive statistics comparing growth of countries that carried out reforms and those that did not. A common pattern is that economic performance was worse for reform adopters in the 1980s and 1990s. This pattern may partly reflect the fact that countries which came under pressure to adopt reforms already tended to be worse off, but may also reflect that such reforms required adjustments that caused short-term hardship to low-income populations that were already struggling. Between 2000 and 2019, median per capita GDP growth was higher than during the 1980s and 1990s for both reformers and non-reformers. However, the increase in growth was even higher for reform adopters. While it would be imprudent to draw definitive conclusions from these simple descriptive analyses, the results are consistent with a reversal of the economic fortunes of reform adopters in the last two decades following their initial dismal economic performance during the 1980s and 1990s.

We next explore the role of two alternative explanations for improved growth across the sub-Saharan Africa region since 2000: whether countries received debt relief and the "super-cycle" increase in commodity prices early in that time

period. We find that the post-2000 per capita GDP growth was higher for non-commodity-dependent countries, compared with commodity-dependent countries. Additionally, among the commodity-dependent countries, per capita GDP growth was higher for the earlier reformers compared with non-reformers. For debt relief, countries that benefited from debt forgiveness experienced higher per capita GDP growth compared with countries that did not. Among the countries that benefited from debt relief, reformers generally experienced a higher per capita GDP growth.

To enrich the aggregate analysis, we present three case studies for Nigeria, Uganda, and Ethiopia. This discussion illustrates how the factors of economic reform, debt relief, and commodity prices interact, and also emphasizes the potential importance of other factors like national investment in infrastructure. An overall message is that implementing economic reforms successfully requires a stable government and socio-political environment, which in turn requires a focus on the poor and on those negatively affected by reforms to sustain needed public support.

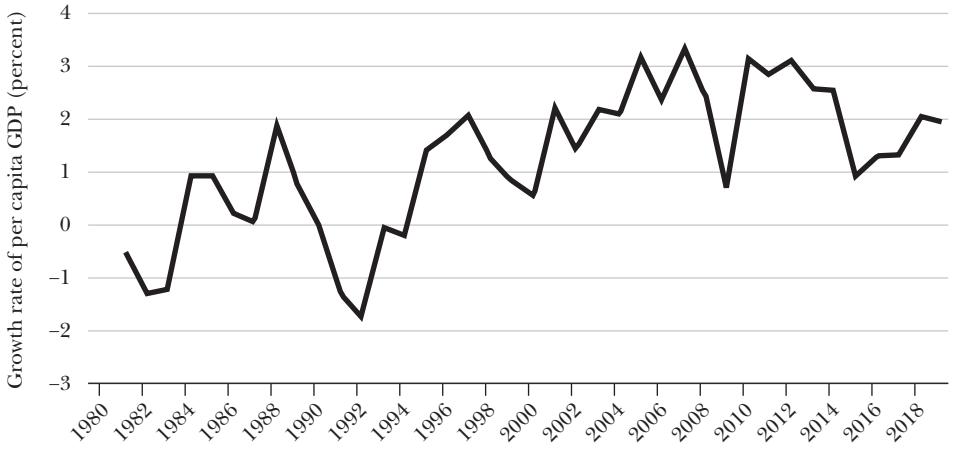
Existing Evidence on Washington Consensus Policies in Sub-Saharan Africa

One can make a prima facie case that something changed for the better with regard to the economies of sub-Saharan Africa in the early 2000s. As shown in Figure 1, African economies have experienced remarkable improvement in economic growth, with median country real GDP per capita growth rising from 0.2 percent per year on average in the 1980s and 1990s to 1.6 percent over 2000 to 2019. Figure 2 shows that the rate of inflation in the region for the median country declined from double digits in the 1980s and 1990s, including a peak of 25 percent inflation for the median country in 1994 (partly caused by the devaluation of the African Financial Community or CFA franc in 1994, as discussed in Franses and Janssens 2018), to stabilize at around 5 percent in the past two decades.

These observations raise the question of whether the market-oriented reforms of the 1980s and 1990s could have played a role in the region's improved economic performance of the past two decades. The hope at that time was that market-oriented reforms would correct domestic policy-induced distortions in prices, such as overvalued exchange rates, subsidies that led to artificially low agricultural commodity prices, high wage rates, low interest rates, and subsidized input prices (Due and Gladwin 1991; Williamson 1993; Easterly 2019; Chari, Henry, and Reyes 2020). Similarly, market-based policies like privatizing public enterprises, removing or relaxing exchange rate controls that biased export trade towards certain commodities, and fiscal adjustment to balance budgets by reducing spending on subsidies would support stronger economic growth.

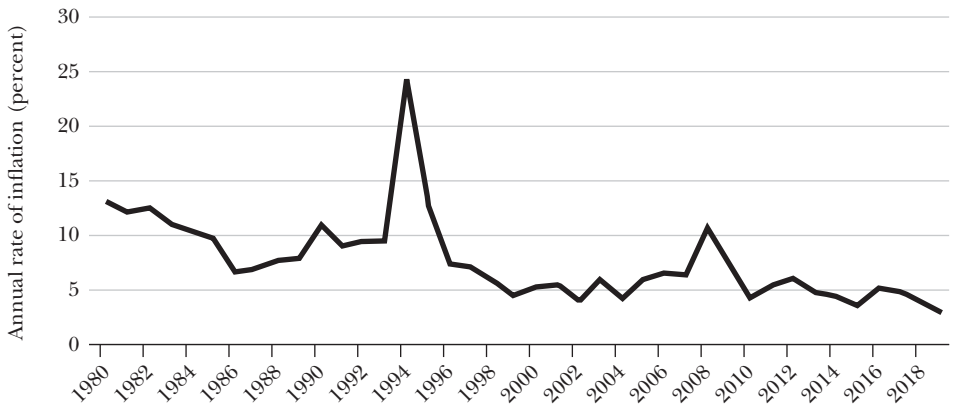
Most of the early literature found that the reform policies failed to improve economic conditions in African countries. Perhaps the most common reason for this outcome centered on the failure of the reforms to account for political economy within countries: in particular, a sense that reforms were being imposed by outside

Figure 1
Median Real GDP Per Capita Growth Rates in Sub-Saharan Africa, 1980–2019



Source: World Bank
Note: Initial reform period between 1980 and 1999.

Figure 2
Median Inflation in Countries of Sub-Saharan Africa, 1980–2019



Source: Consumer Price Index data from the World Bank
Note: Initial reform period between 1980 and 1999.

agents as a condition for debt relief or additional loans without adequately emphasizing the role of local ownership in shaping domestic economic policy (Ekpo 1992; Easterly 2000; Due and Gladwin 1991; Birdsall, Caicedo, and De la Torre 2010; Adedeji 1999; Mkandawire and Olukoshi 1995; Rodrik 2006; Stiglitz 2005). Other studies attributed the failures of the reforms to increases in domestic inflation and

its adverse effect on real incomes and well-being post-reform (Due and Gladwin 1991; Ekpo 1992). The negative effects of the reforms were also disproportionately felt by rural farmers, especially women working in food crop production. Ironically, while international financial institutions were advocating for the removal of agricultural subsidies in Africa, the advanced economies, including the United States and other high-income countries, heavily subsidized agricultural production, making it difficult for African farmers to compete (Due and Gladwin 1991; Mkandawire and Olukoshi 1995). Thus, these market-oriented reforms increased unemployment and socio-political unrest in several African countries in the 1980s and 1990s (Mkandawire and Olukoshi 1995; Due and Gladwin 1991; Ihonvbere 1993; Elson 1995).

However, a more recent literature has suggested that the reforms were successful in improving economic growth over time, particularly when policymakers had the state capacity to implement them (Prati, Onorato, and Papageorgiou 2013; Grier and Grier 2020; Dollar and Svensson 2000). Conversely, these studies suggest that the de facto reductions in state capacity required by some reforms may have contributed to their failure in some countries. For instance, the ratio of civil servants to the population in sub-Saharan Africa as a whole fell to 1 percent in 1996, lower than the 3 percent for other developing countries and much lower than the OECD average of 7 percent (Sender 1999). Without a motivated, well-equipped, civil service, proper implementation and regulation of these reforms was often incredibly difficult.

Descriptive Evidence on the Effects of Reforms

For the purpose of this discussion, we classify the ten Washington Consensus policies (as discussed in Serra and Stiglitz 2008), into three main categories: 1) fiscal policy reform, which includes fiscal discipline, reordering of public expenditures toward pro-poor priorities, tax reforms to broaden the base and hold down marginal tax rates; 2) domestic market-oriented reforms, which includes interest rate liberalization, privatization, deregulation to reduce barriers entry and exit of firms, and legal security for property rights (especially in the informal sector); and 3) openness reforms, which include liberalization of inward foreign direct investment, trade liberalization, and competitive exchange rates.

For each of the three main categories, we choose one indicator to represent changes in this area. For fiscal policy reform, we treat a country as a fiscal reformer if the average primary fiscal balance in 1995–1999 is higher than –0.7 percent of GDP, which is the median for the region. For domestic market reform, we treat a country as a reformer if the cumulative number of privatizations deals from 1988 to 1999 is greater than or equal to six, which is the median for the region. Finally, for openness, we treat a country as a reformer if it was open to trade, as defined in Sachs and Warner (1995), for at least five years from 1980 to 1999. They classify a country as "open" to trade if it does not have any of the following: average tariff rates in excess of 40 percent; non-tariff barriers that cover more than 40 percent of imports;

Table 1
Reform Countries in Sub-Saharan Africa

Domestic market-oriented reforms	*Trade openness*	*Fiscal reforms*
Burundi	Benin	Angola
Benin	Botswana	Benin
Cote d'Ivoire	Cote d'Ivoire	Central African Republic
Ghana	Cameroon	Cote d'Ivoire
Kenya	Cabo Verde	Congo, Dem. Rep.
Mozambique	Ghana	Congo, Rep.
Malawi	Guinea	Gabon
Nigeria	Gambia, The	Guinea
Senegal	Guinea-Bissau	Guinea-Bissau
Togo	Kenya	Kenya
Tanzania	Mali	Madagascar
Uganda	Mozambique	Mozambique
South Africa	Mauritius	Nigeria
Zambia	Niger	Rwanda
Zimbabwe	Tanzania	Senegal
	Uganda	Eswatini
	South Africa	Seychelles
		Tanzania
		Uganda

Note: See text for details. For fiscal policy reform, we treat a country as a fiscal reformer if the average primary fiscal balance in 1995–1999 is higher than –0.7 percent of GDP, which is the median for the region. For domestic market reform, we treat a country as a reformer if the cumulative number of privatizations deals from 1988 to 1999 is greater than or equal to six, which is the median for the region. Finally, for openness we treat a country as a reformer if it was open to trade, as defined in Sachs and Warner (1995), for at least five years from 1980 to 1999. Some non-reformers not included in the above list include Namibia, Sudan, Eritrea, Somalia, and Sierra Leone.

a socialist economic system; the state has a monopoly on major exports; and a black market currency-trading premium in excess of 20 percent.

Table 1 shows the classification of countries by reform categories. Of the 32 sub-Saharan African countries for which we have data, 59 percent are fiscal reformers. Of the 36 countries for which we have data, 42 percent are domestic market reformers and 47 percent are openness reformers. Six countries—Benin, Cote d'Ivoire, Kenya, Mozambique, Tanzania, and Uganda—are reformers in all three categories. There is admittedly some subjectivity in these rules. After all, whether a nation is a "reformer" is not a binary yes-or-no question. The classification used here mostly distinguishes between those who enacted the most reforms and those who reformed the least. Still, this descriptive approach provides useful insights.

Table 2 summarizes the results showing the trends in per capita GDP growth rates for reformers and non-reformers for all countries and by reform category. Overall, the median GDP per capita growth was slightly positive (0.2 percent) across the region between 1980 and 1999. From 2000 to 2019, the growth rate rose

Table 2
Reforms and Changes in Median Per Capita Real GDP Growth (%)

Reforms	*Type*	*1980–1999*	*2000–2019*	*Difference*
	All countries	0.2	1.6	+1.4
Fiscal reforms	Reformers	–0.3	1.5	+1.8
	Non-reformers	1	1.8	+0.8
Domestic market-oriented reforms	Reformers	–0.6	1.5	+2.1
	Non-reformers	0.2	1.6	+1.4
Trade-openness	Reformers	0.8	1.9	+1.1
	Non-reformers	–0.2	1.1	+1.3

Note: See text for details. For fiscal policy reform, we treat a country as a fiscal reformer if the average primary fiscal balance in 1995–1999 is higher than –0.7 percent of GDP, which is the median for the region. For domestic market reform, we treat a country as a reformer if the cumulative number of privatizations deals from 1988 to 1999 is greater than or equal to six, which is the median for the region. Finally, for openness we treat a country as a reformer if it was open to trade, as defined in Sachs and Warner (1995), for at least five years from 1980 to 1999. GDP growth rate data based on median per capita real GDP growth rates across groups and time periods. Median annual growth in constant per capita GDP figures from World Bank data.

by 1.4 percentage points to 1.6 percent. However, performance varied by reform categories.

The premise of the Washington Consensus policies reforms rested on two interdependent and testable hypotheses: first, in the years following the reforms, economies that adopted reforms would perform better than they did in the preceding years and, second, reform adopters would outperform non-reformers. Here, we examine the links between reform adoption and the region's economic performance, as measured by per capita GDP growth.

Across sub-Saharan Africa, the median budget deficit declined from –2 percent of GDP in the early 1980s to –0.7 percent of GDP in the late 1990s, suggesting an increase in fiscal discipline across the region. The reduction in the deficit continued through 2010: indeed, budget deficits for the region as a whole were near-zero from 2005 to 2009. However, deficits widened afterwards due partly to the effects of the global financial crisis of 2008–09 and a substantial terms-of-trade shock in 2014.

Africa's fiscal reformers in the 1980–1999 period experienced negative growth rates, with the trends reversing sharply in the post-2000s era. Comparing the two sets of countries as shown in Table 2, the per capita GDP growth rate was slightly higher for non-reformers than for reformers over the past two decades. However, the per capita GDP growth rate increased by more for reformers compared with non-reformers between 1980 and 1999 and 2000 and 2019, consistent with, but not conclusive confirmation of, the positive long-run predictions of reform adoption for economic performance.

We use privatization as a proxy for domestic market-oriented reforms (Parker and Kirkpatrick 2005), in part because no comprehensive and reliable cross-country measures were available for other Washington Consensus goals like deregulation, legal security for property rights, and interest rate liberalization. Moreover, privatization is often regarded by both supporters and opponents of the Washington Consensus reforms to be a key feature of domestic policy. In sub-Saharan Africa, the share of countries with at least one privatization deal rose from 5 percent in 1988 to reach 40 percent in the late 1990s. Similarly, the number of enterprises privatized increased significantly from just three in 1988 to 160 in 1996. The pace of privatization varied across the region. While some countries, including Cote d'Ivoire, Uganda, Nigeria, Ghana, Kenya, Mozambique, Zambia, and Tanzania, privatized more than 50 state-owned enterprises between 1988 and 1999, others, including Gabon, Mauritius, Rwanda, Equatorial Guinea, Gabon, and Central African Republic, did not privatize any over this period. In some cases, the push to privatize state-owned enterprises was part of a strategy to consolidate fiscal balances.

Previous scholars have highlighted the serious challenges faced by African countries in the planning and implementation of privatization policies while at the same time pursuing other aspects of structural adjustment plans and debt-relief negotiations in environments of incomplete markets and weak enforcement capacity (Bayliss and Cramer 2003; Ariyo and Jerome 1999). Although the Washington Consensus framework did recognize the importance of complete markets and proper regulation as preconditions for successful privatization, these caveats were often overlooked in policy design. In particular, international financial institutions often failed to highlight adequately that privatization reforms should be accompanied by antitrust legislation in promoting competitive markets. They also underestimated the effects of rapid privatization on the morale of public sector employees, who were essential for proper regulation of the privatization process (Bayliss and Cramer 2003; Ariyo and Jerome 1999).

Table 2 shows the average performance between those countries with more and fewer privatizations as a proxy for more general domestic market-oriented reforms. Similar to the results of the fiscal reforms, market reformers experienced declines in per capita GDP growth over the reform period from 1980 to 1999, followed by a sharp reversal post 2000. Over 2000–2019, growth rates for reformers and non-reformers looked very similar at 1.5 percent and 1.6 percent on average, respectively. However, the set of countries that privatized the most in the late 1980s and in the 1990s experienced a much higher increase in median real GDP per capita growth in the last two decades: 2.1 percentage points compared with 1.4 percentage points for the non-reformers.

Finally, sub-Saharan Africa increasingly opened to trade in the 1980s and 1990s. In the early 1980s, only 5 percent of the countries were classified as open to trade. That share rose to reach almost 60 percent by 2000. Around the same period, African countries accelerated the adoption of more competitive exchange rates: for example, the share of countries with floating or semi-floating exchange

rates rose from 45 percent in 1980 to 60 percent in the early 1990s. While trade openness increased, previous scholars have highlighted that this did not translate to immediate increases in investment in sub-Saharan Africa. Indeed, cuts in public investment to adhere to fiscal reforms contributed to the decline in investment (Sender 1999).

Many trade liberalization policy reforms undertaken over this period underestimated the role of incentives facing producers in incomplete markets. Liberalization in the agricultural sector, hastily implemented, negatively impacted terms of trade for farmers who were sometimes unable to compete with international prices (Sender 1999). Higher prices for agricultural commodities in the 1980s and 1990s worsened local food shortages and led to protests in African countries (Herbst 1990). Indeed, these events may have also contributed to the steep reductions in Africa's aggregate investment levels in the early 1980s.

Despite initial reductions in total investments across the continent in the early part of the reform years, countries that adopted trade openness reforms experienced small positive growth rates over 1980–1999. Real GDP growth per capita increased for both reformers and non-reformers between 1980 and 1999 and between 2000 and 2019. The increase was roughly comparable for reformers and non-reformers, although reform countries ended up with higher growth rates of 1.9 percent in the 2000–2019 period.

Taken as a whole, this descriptive evidence is consistent with the earlier work: that is, reforming countries tended to be worse performers before 1990 but made a more substantial jump in growth rates after 2000.

Possible Alternative Explanations for Post-2000 Economic Performance

In this section, we assess two plausible alternative, and not necessarily mutually exclusive, explanations for the improved economic performance of sub-Saharan Africa over the past two decades. One explanation is that African countries benefited from debt relief and the resulting additional fiscal space allowed governments to increase public expenditures to boost economic growth. A second explanation is that African countries benefited from the sustained increase in commodity prices in the early 2000s, driven, in part, by the high demand from China, and known as the commodity super-cycle (Fernández, Schmitt-Grohé, and Uribe 2020).

Debt Forgiveness

Beginning in the 1990s, officials from major creditor countries (a group known as the Paris Club) and multilateral organizations adopted the ambitious Multilateral Debt Relief Initiative for outright forgiveness of debt owed by a group of 36 low-income countries—29 of them located in Africa. This debt relief effort was the logical advancement of a variety of initiatives for debt relief, the most prominent of which was the Heavily Indebted Poor Countries (HIPC) initiative instituted by the IMF and

World Bank in 1996 to address debt overhang in the poorest countries of the world. A list of African countries scheduled for debt relief under the HIPC program is shown in Table 3. A total of 32 countries, 67 percent of countries in sub-Saharan Africa, were classified as HIPC countries, accounting for a total of $239 billion in (constant 2010) GDP in 2000. In contrast, the total GDP in 2000 for the 16 non-HIPC countries listed in Table 3 was higher at about $560 billion (by World Bank estimates).

The debt relief initiatives were expected to improve economic performance. After unloading the inherited debt overhang, an infusion of new loans, improved policies, and enhanced investment incentives were expected to increase economic and social development outcomes. Some previous evidence has shown positive correlations between reduced debt burdens and economic upturns (Coulibaly, Gandhi, and Senbet 2019). The average public debt level (as a percentage of GDP) for sub-Saharan Africa declined to about 36 percent in 2012 from highs of around 110 percent in 2001, significantly below the levels leading up to the HIPC initiative.

Table 4 offers a comparison of African countries that benefited from debt relief and those that did not. Countries receiving debt relief might be expected to be in worse overall economic shape at the start of the process and, indeed, the growth rate for beneficiaries of debt relief was lower between 1980 and 1999 than non-debt relief recipients. However, growth rates were similar between debt relief and non-debt relief countries at 2 percent over 2000–2019. Thus, countries that received debt relief experienced higher increases in per capita economic growth over the last two decades, 2.3 percentage points, compared with 0.3 percentage points for the countries that did not receive debt relief.

Debt relief through programs was often conditioned on strict adoption of market liberalization reforms like those outlined in Washington Consensus policies. Indeed, many of the reforms undertaken by African countries in the 1990s were initiated with an objective to reach certain debt relief eligibility targets (Ekpo 1992; Sender 1999) and hence, there was significant overlap between reform adopters in Table 1 and debt relief recipient countries in Table 3. A full and persuasive decomposition of patterns and trends between reform adopters and the debt relief recipients would be a challenging task. But this descriptive comparison of patterns and trends over this period shows that, among the beneficiaries of debt relief, the countries that adopted fiscal and market-oriented reforms posted higher economic growth than non-adopters. However, there do not appear to be significant differences in growth rates between the 1980–1999 reform period and the post-2000s era for debt relief recipients that adopted more trade-openness reforms.

The Commodity Super-Cycle

Commodities have featured heavily in the exports of many African countries for a number of years, with exports of commodities like oil as high as over 80 percent of total exports in countries like Angola, Congo, and Nigeria in 1990 and through the early 2000s (Deaton 1999). Minerals like diamonds and uranium have also featured heavily in commodity exports of African economies like Botswana (where

Table 3

African Countries by Debt Relief under Heavily Indebted Poor Country (HIPC) Program and Commodity-Dependent Status

HIPC countries	*Non HIPC countries*	*Commodity-dependent countries*	*Non-commodity-dependent countries*
Benin	Angola	Benin	Cape Verde
Burkina Faso	Botswana	Burkina Faso	Comoros
Burundi	Cape Verde	Burundi	Djibouti
Cameroon	Djibouti	Cameroon	Kenya
Central African Republic	Equatorial Guinea	Central African Republic	Lesotho
Chad	Gabon	Chad	Liberia
Comoros	Kenya	Congo, Dem. Rep.	Madagascar
Congo, Dem. Rep.	Lesotho	Eritrea	Mauritius
Eritrea	Mauritius	Ethiopia	Niger
Ethiopia	Namibia	Gambia	São Tomé and Príncipe
Gambia	Nigeria	Ghana	Senegal
Ghana	Seychelles	Guinea	South Africa
Guinea	South Africa	Guinea Bissau	Swaziland
Guinea Bissau	South Sudan	Ivory Coast	Togo
Ivory Coast	Swaziland	Malawi	Uganda
Liberia	Zimbabwe	Mali	
Madagascar		Mozambique	
Malawi		Congo, Rep.	
Mali		Rwanda	
Mozambique		Sierra Leone	
Niger		Somalia	
Congo, Rep.		Sudan	
Rwanda		Tanzania	
São Tomé and Príncipe		Zambia	
Senegal		Angola	
Sierra Leone		Botswana	
Somalia		Equatorial Guinea	
Sudan		Gabon	
Togo		Namibia	
Uganda		Nigeria	
Tanzania		Seychelles	
Zambia		South Sudan	
		Zimbabwe	

Note: See text for details. Debt relief countries are HIPC countries as classified by the World Bank. Commodity-dependent countries are as classified by the IMF and defined as countries where commodities account for ≥ 80% of merchandise exports. The designation of HIPC and commodity-dependent categories is using 2016 data. While the categories change over time, there is a strong positive correlation between HIPC and commodity-dependent designation in the 1980s/1990s and as of the most recent data we use here, so the categories using the most recent data available are informative for our study (Djimeu 2018).

diamonds were 80 percent of exports in 1990) and Niger (where uranium was 83 percent of exports in 1990). In the 2000s, commodity prices surged in response to higher demand from emerging market economies, notably China, as well as from concerns over long-term supply. A notable example was the boom in oil prices over this period, with oil prices rising over 200 percent from $30 per barrel in 2000 to $100 per barrel in 2008.

Table 4

Reforms and Changes in Median Per Capita Real GDP Growth (%) by Debt Relief Recipient Status

Type	*Reforms*	*1980–1999*	*2000–2019*	*Difference*
All countries		0.2	1.6	+1.4
Non debt relief		1.7	2	+0.3
Debt relief	All beneficiaries of debt relief	-0.3	2	+2.3
	Fiscal reformer	–0.4	2.2	+2.6
	Fiscal non-reformer	0.4	1.6	+1.2
	Market reformer	–0.1	2.5	+2.5
	Market non-reformer	–0.3	1.8	+2.1
	Openness reformer	–0.2	2.2	+2.3
	Openness non-reformer	–0.7	1.8	+2.6

Note: See text for details. Beneficiaries of debt relief refer to HIPC countries. GDP growth rate data based on median per capita real GDP growth rates across groups and time periods. Median annual growth in constant per capita GDP figures from World Bank data.

The commodity price super-cycle was then disrupted during the 2008–2009 global financial crisis and, subsequently, by an adverse terms of trade shock for Africa's exporters in 2014. Despite these shocks, higher commodity prices over much of the past two decades benefited several commodity-dependent countries. We define commodity dependence according the IMF definition of countries where commodities account for more than 80 percent of total merchandise exports. As shown in Table 3, 33 countries, 69 percent of countries in sub-Saharan Africa, were classified as commodity-dependent countries, accounting for a total of $452 billion in (constant 2010) GDP in 2000. In contrast, the total GDP in 2000 for the 15 non-commodity-dependent countries in sub-Saharan Africa listed in Table 3 was lower at about $347 billion (based on World Bank estimates).

Table 5 shows a breakdown of growth rates for both commodity-dependent and non-commodity-dependent countries. Both groups experienced higher GDP per capita growth between 2000 and 2019 compared to the previous two decades. In fact, the increase in per capita GDP growth rate was higher for non-commodity-dependent countries, 1.9 percentage points compared with 1.4 percentage points for commodity-dependent countries.

This pattern seems to suggest that although the commodity price super-cycle likely played an important role from 2000 to 2006, when comparing the longer periods as in Table 5, its differential effect on longer-term growth of African countries is not substantial. Indeed, per capita GDP growth averaged 2 percent between 2000 and 2004 before commodity prices began their rapid ascent (Coulibaly 2017), suggesting that the increase in commodity prices was not the sole driver of the post-2000 economic performance for sub-Saharan Africa. As shown in Table 5, an examination of the trends in growth rates among commodity-dependent countries

Table 5

Reforms and Changes in Median Per Capita Real GDP Growth (%) by Commodity-Dependent Status

Type	*Reforms*	*1980–1999*	*2000–2019*	*Difference*
All countries		0.2	1.6	+1.4
Non-commodity-dependent		0.4	2.2	+1.9
Commodity-dependent	All dependent countries	0.4	1.8	+1.4
	Fiscal reformer	–0.1	1.5	+1.6
	Fiscal non-reformer	0	1.6	+1.7
	Market reformer	0	2.5	+2.5
	Market non-reformer	0.1	1.5	+1.4
	Openness reformer	0.5	1.9	+1.4
	Openness non-reformer	0.2	1.3	+1.1

Note: See text for details. Commodity-dependent countries based on IMF data classifications. GDP growth rate data based on median per capita real GDP growth rates across groups and time periods. Median annual growth in constant per capita GDP figures from World Bank data.

between reformers and non-reformers shows that countries that adopted market reforms like privatization and trade openness posted higher increases in median growth rates between 1980 and 1999 and 2000 and 2019. For fiscal reformers, in contrast, there appears to be no discernible difference in growth rates.

The results suggest that debt relief may have also contributed to the higher per capita economic growth of the last two decades, with less of an effect for commodity prices. Within the categories of countries that were beneficiaries of debt relief and commodity dependent, reformers generally posted larger growth gains between the reform period and the post-2000s era, suggesting that reforms may have played a role in improving economic performance, independently of the commodity price boom and debt relief.

Select Country Experiences

The analysis so far has taken a broad-brush approach to examining the links between Washington consensus policy adoption and economic performance in Africa. To complement and enrich the discussion on the regional experience, we explore the reform experience in three countries with different situations, implementation approaches, and results, featuring two countries with the largest populations in Africa (Nigeria and Ethiopia) and what is widely viewed as a case of successful reform adoption (Uganda). The case studies also represent two reform countries (Nigeria and Uganda) and one non-reform country (Ethiopia), if categorized according to the domestic market-oriented, trade openness and fiscal reforms classifications discussed in the previous sections.

Nigeria

Nigeria scored highly on both domestic market-oriented reforms and fiscal reforms (as shown in Table 1) and is a commodity-dependent country that was not one of the countries scheduled for debt relief (as shown in Table 3). Nigeria has been heavily dependent on oil exports since the 1970s. In the 2000s, over 70 percent of Nigeria's government revenue comes from petroleum, with petroleum exports as a share of total exports growing to over 90 percent in the 2000s (Archibong 2018). The heavy dependence on oil exports has made the country very vulnerable to external price shocks, with deleterious implications for the ability to finance public spending and debt (Okonjo-Iweala 2014). Swings in oil prices played a major role in creating Nigeria's debt problem in the 1980s, but after 2000, a combination of improved management of oil resources and improved macroeconomic policies helped to improve Nigeria's growth.

Global oil prices crashed (in nominal terms) from about $30 per barrel in the early 1980s to about $12 per barrel in the mid-1980s, significantly increasing Nigeria's debt-to-GDP ratio. Under pressure to reach agreements on debt rescheduling, Nigeria implemented policy reform in the form of structural adjustment programs with the support of the IMF and World Bank (Ekpo 1992; Devarajan, Dollar, and Holmgren 2002). Previous work has described the Nigerian economic experience post-policy adoption in the 1980s as dismal by citing decreases in GDP growth rates from 6.9 percent pre-adjustment to –1.7 percent in the postperiod (Ekpo 1992), but this also compares the period of high oil prices to the period after oil prices crashed.

Nigeria's reforms focused on fiscal tightening and privatization (as shown earlier in Table 2), but also induced severe cuts in social spending on education and health, which led to increased hostility for the reforms in the 1980s and 1990s by Nigerian citizens. (In contrast, the list of actual Washington Consensus policies in Williamson (1993) explicitly emphasizes reorientation of spending toward pro-poor programs.) Nigeria's reforms were then abandoned by the Babangida military regime and the country continued to be beset by poor macroeconomic policy. Nigeria continued to borrow and accumulated up to $30 billion in debt to the Paris Club of creditors even though the country earned more than $300 billion in crude oil revenues over the 1970s–2001 period (Okonjo-Iweala 2014). While some of the oil revenue and borrowed money was invested in needed infrastructure, education, and health, lack of monitoring of spending and opaque ad hoc budgets meant there was a significant amount of spending on "white elephant" projects like unproductive steel mills.

Following the transition to democracy in 1999 and under the helm of then-President Olusegun Obasanjo, Nigeria was faced with an unstable macroeconomic environment in the early 2000s: volatile exchange rates, double-digit inflation (23 percent per year in 2003), a relatively high fiscal deficit (3.5 percent of GDP in 2003) and low GDP growth (2.3 percent on average for the previous decade). The country embarked on macroeconomic reforms (under then-finance minister Ngozi Okonjo-Iweala), again with a focus on privatization and budget monitoring, but this

time also with a notable investment in education and health. In addition, to reduce volatility in public finances, Nigeria adopted an oil price-based fiscal rule that used the long-run, 10-year average oil price to set government budgets and targets for spending. Based on the rule, when oil prices were above average, the government would set aside some excess revenues from oil in the form of a savings account called the Excess Crude Oil Account. The fiscal rule, which was institutionalized in national law in the Fiscal Responsibility Act signed in 2007, linked savings to fiscal discipline around government spending, aiming for a fiscal deficit of 3 percent of GDP. The Excess Crude Oil Account policy was successful both in building fiscal discipline and helping Nigeria weather shocks like the financial crisis of 2008–2010, when oil prices fell from over $140 to $40 per barrel. Over this period, Nigeria was able to draw on savings from the account to implement a fiscal stimulus of around 0.5 percent of GDP and to maintain public spending.

Increased public savings between 2004 and 2006 as a result of policy led to fiscal surpluses of 7.7 percent of GDP in 2004 and 10 percent of GDP in 2005. This laid the groundwork for a relief of a $30 billion debt, of which $18 billion was completely written off by the Paris Club, and Nigeria paid off its external debt arrears of about $6 billion. In addition, Nigeria was able to increase its foreign reserves from $7 billion in 2003 to $46 billion by the end of 2006, while also implementing tighter monetary policy to reduce inflation from 21.8 percent in 2003 to 10 percent in 2004. These changes also helped to spur private sector investment. Growth averaged 8.1 percent a year from 2003 to 2006, and the share of spending on health and education rose to 5 percent and almost 10 percent for health and education in 2007, respectively.

Reforms in the early 2000s also targeted sectors that were large drains on public finances for privatization, including the telecommunications sector, the downstream petroleum sector, and the power sector, with varying degrees of success. Nigeria also benefited from the increase in oil prices in the post-2000 period, and both the reforms and increases in prices combined to create an attractive environment for private investors in the country. The Nigerian experience with reforms, and specifically the contrast between the outcomes of reforms under the military versus democratic regimes mentioned here, highlights the importance of a committed government centering social welfare with pro-poor spending in implementing successful reforms.

Uganda

Uganda is often touted by international financial institutions as an example of successful application of reforms, but digging into the details of reform presents a more mixed picture (Dijkstra and Van Donge 2001; Devarajan, Dollar, and Holmgren 2002; Hickey 2013; Rwamigisa et al. 2018). Of the three key areas of reform discussed earlier, Uganda was one of a handful of countries (six of them listed in Table 1) that scored highly on all three: domestic market-oriented reforms around privatization, fiscal reforms aimed at improving the fiscal balance, and increased trade openness over the 1980–1999 period. Uganda was not one of the

commodity-dependent countries but was one of the countries scheduled for debt relief under the Highly Indebted Poor Countries (HIPC) initiative (as listed in Table 3).

Between 1971 and 1986, Uganda experienced economic decline, but GDP per capita rose by almost 40 percent in the first decade after longtime/current president, Yoweri Museveni, was in power between 1986 and 1996. In 1987, the country received an IMF loan, with loan renewals occurring from 1989 to 1992 and again from 1992 to 1997. Real GDP per capita grew on average 4.2 percent per year between 1992 and 1997. The two main reforms mandated by the IMF in Uganda were trade liberalization and the progressive reduction of export taxation; in Uganda, coffee was the main export crop. The benefits of liberalized cash crop exports were large but also limited and unequally distributed, with only a small number of rural coffee farmers experiencing increases in rural per capita incomes over the period of policy reform from 1988 to 1995.

Uganda also privatized a substantial number of public enterprises, including industries in banking, insurance, railways, and telecommunications—a set of moves that was highly criticized within the country. The main critique was that the privatization had proceeded too rapidly, with relatively little oversight. As a result, the privatizations benefited government and corporate interests of advanced economies rather than the Ugandan population. While public spending in healthcare increased, it did not keep pace with government spending, so that the share of health in the budget declined slightly between 1989 and 1994. In 1998, Uganda was also the first country to receive debt relief under the HIPC initiative, some $650 million reduction in Uganda's multilateral debt stock, but then the reduction was delayed by a year, which amounted to $193 million in lost relief benefits. With the delay, public funds were diverted from spending on healthcare provision toward debt repayments. A key difference between the Nigerian and Ugandan cases at this stage was the relatively higher commitment and spearheading of reform policies in Uganda. The strong commitment from Uganda could have been due to the country's highly indebted/HIPC status and high level of external financing as well, which was accompanied by critiques about donor pressure in spearheading reforms (Hickey 2013).

Between 2000 and 2019, Uganda, a non-commodity-dependent country, experienced stable growth rates of around 6.3 percent per year on average. Reforms in the agricultural sector have been credited with halving between 1992 and 2013 the share of households in poverty. The details of the success of some of the agricultural sector reforms have also come under criticism in recent studies, with a prominent example being the National Agricultural Advisory Services (NAADS) program, first implemented in 2001 (Rwamigisa et al. 2018). The NAADS reform was aimed at increasing market-oriented agricultural production by "empowering farmers to demand and control agricultural advisory services," which included replacing public sector extension agents with contracted private service providers (Rwamigisa et al. 2018). Although early evidence from the program heralded the program's success in 2007, particularly in encouraging farmer adoption of new crops and agricultural

production technologies and practices, more recent evidence has found more mixed results on the program's success, with studies citing mismanagement of public funds and low technological uptake by farmers as obstacles (Benin et al. 2007; Rwamigisa et al. 2018). The program was eventually scrapped in 2014, with agricultural extension services duties transferred back to the Ministry of Agriculture, Animal Industry and Fisheries. Despite these reforms, Uganda still faces challenges in translating recorded growth rates into improvements in human capital, like reductions in child stunting and increases in educational attainment for most of its population.

Ethiopia

Ethiopia is the second most-populous country in sub-Saharan Africa (after Nigeria). Ethiopia's experience has not had much success as a reformer, and it does not score highly or feature as a reform adopter on any of the three classifications discussed previously. It was also a country that was both commodity dependent and listed for debt relief as one of the Highly Indebted Poor Countries (HIPC) listed (shown in Table 3).

In the 1980s, the country was immersed in a civil war under the military regime the Derg and struggled to implement reform during significant political and economic crises (as described in Devarajan, Dollar, and Holmgren 2002). Economic policy under the Derg was notorious for granting monopolies to the state over imports and exports, with high tariffs and heavy investment in the public sector (Oqubay 2018). Towards the end of the Derg era and with the introduction of the new communist government, the People's Democratic Republic of Ethiopia in 1987, the government adopted a few financial stabilization policies in the early 1990s, including infrastructure investment as well. However, with weak state capacity, promoting development and financial stabilization amidst a civil war made attempts at reform an arduous process.

After the end of the civil war, and following the dissolution of the People's Democratic Republic of Ethiopia in 1991, a coalition of political parties under the Ethiopian People's Revolutionary Democratic Front (EPRDF) took over the country from 1991 till 2019. The EPRDF government explicitly pursued industrial policy with active government involvement in agriculture as the assumed key for economic growth between 1995 and 2015 (Oqubay 2018). An example of this was the government's adoption of the Agricultural Development Led Industrialization (ADLI) strategy in 1994, which it then proceeded to follow for over two decades, with a focus on investment in agriculture. While the government received substantial debt relief and official development assistance from donors like the IMF, particularly in the early part of the regime (the ratio of official development assistance to gross national product rose from 12 percent in the 1980s to 23 percent in the 1990s), it did not adopt many of the reforms proposed under the Washington Consensus, choosing instead a so-called "gradualist" approach that involved a mixture of some liberalization like privatization of a few state-owned enterprises in specific sectors (for example, banking, utilities. and air travel) along with industrial policy (Tekeste 2014; Oqubay 2018; Abegaz 1999). Other sectors like retail businesses along with some

banking and domestic freight services were closed to foreign investment and open only to Ethiopians—a fact which was sometimes a point of contention with lending institutions like the IMF (Oqubay 2018). Ethiopia saw significant increases in growth over this period in the 1990s, with average annual real GDP growth increasing from 3 percent in 1990–91 to 7.8 percent between 1995 and 97, and inflation rates falling from 21 percent in 1991–92 to 3.6 percent in 1993–98 (Abegaz 1999).

Among policy instruments used were industrial financing, including investment financing through the Development Bank of Ethiopia and Commercial Bank of Ethiopia, export promotion through target setting, retention of foreign exchange earnings, and exchange rate policies like devaluation and allocation of foreign exchange to certain sectors. Other policy instruments implemented by the EPRDF government include import tariffs, some privatization of state-owned enterprises in specific sectors, and investment support towards the horticulture and cement industries (Oqubay 2018). In the early 2000s, around half of the federal government's budget for its consecutive five-year programs was designated for pro-poor and high growth sectors (Oqubay 2018). Since 2015, Ethiopia's government has also focused on investment in the manufacturing sector as a key for economic development.

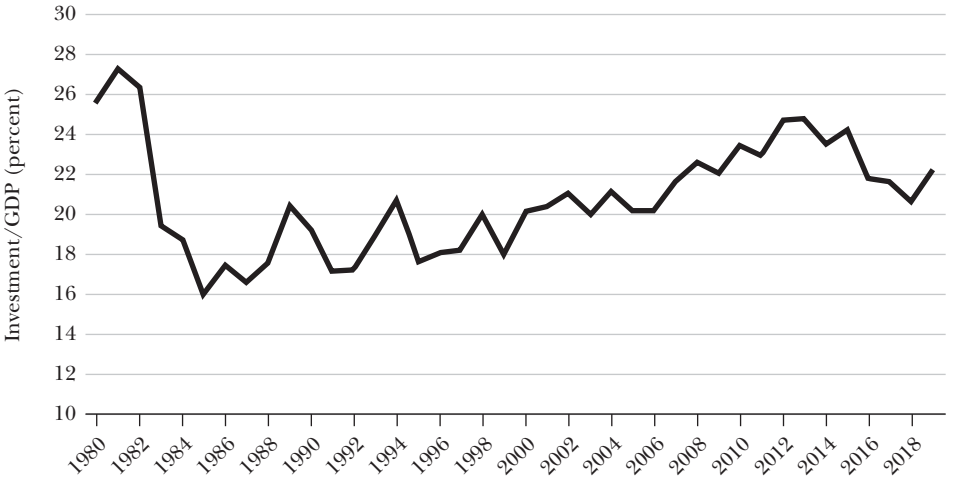
Despite not being one of the reform adopters, and explicitly pursuing industrial policy with active government involvement, Ethiopia has consistently ranked among the top economic performers in the region for much of the past decade and a half, with an average growth rate of real GDP of 8.9 percent between 2000 and 2019. Much of this growth has been attributed to public investment in key infrastructure along with interventions in the agricultural sector to improve productivity and facilitate structural transformation. There has also been a reallocation of labor from low productivity agriculture to more productive industrial and service sectors in the country.

Discussion and Concluding Remarks

Growth in the countries of sub-Saharan Africa improved substantially after around 2000. To what extent can the "Washington Consensus" reforms claim a share of the credit? The descriptive evidence in this paper certainly does not establish a causal effect of the Washington Consensus policies on economic performance in Africa. In addition, as illustrated through the country case studies, reform experience and outcome differed across countries depending on the particular policy and macroeconomic environments, along with the specific policy objectives of governments in different countries.

That said, the reforms related to the Washington Consensus in a number of cases did lead to an improved macroeconomic environment with lower inflation combined with debt reductions. These changes did help to attract more private investment in key sectors like retail, wholesale, telecommunications, and manufacturing that accounted for a significant share of the growth increases in the 2000–2019 period. In addition, the countries of sub-Saharan Africa saw a wave of

Figure 3
Total Investment as a Percentage of GDP in Sub-Saharan Africa, 1980–2019



Source: IMF and UNCTAD

democratization in the 1990s, with the number of countries that held multi-party elections increasing from just two (Botswana and Mauritius) before 1989 to 44 of 48 countries—or 92 percent of sub-Saharan Africa—by mid-2003 (Lynch and Crawford 2011). This had the effect of encouraging investment in infrastructure and in pro-poor policies. While total investment as a share of GDP in sub-Saharan Africa fell sharply in the early 1980s as shown in Figure 3, investment stabilized and then rose. Over time, countries that were more responsive to their citizens as well as international financial institutions learned from past experiences and improved design and implementation of reforms.

As this emphasis on democratization helps to make apparent, we believe that the story of Africa's growth surge in the last two decades also relies heavily on a number of factors that go beyond the economic reform packages of the 1980s and 1990s. Here are some of the reasons why. First, many of the especially indebted countries that came under pressure to carry out reforms were already suffering from lower per capita economic growth over much of the reform period from 1980 to 1999. Thus, comparing the experience of reformers and non-reformers involves some selection bias: the low economic performance may have been a motivator for the reforms, or the lower economic performance may have resulted from the short-term negative effect of the reforms.

Second, we believe that the speed with which many of these reforms were carried out initially, especially domestic reforms like privatization of state-owned enterprises, without careful consideration of the environment of incomplete markets and the institutional challenges faced by African governments, affected the initial effectiveness of

policy implementation and contributed to lower growth rates during the 1980–1999 reform period. Indeed, a difficulty in judging the Washington Consensus framework is that the form spelled out by Williamson (1993) contained a number of conditions that were often lost in policy design. For example, the framework advocated for pro-poor fiscal expenditures and advised against abolishing deregulation designed for safety or environmental reasons. It cautioned against capital account liberalization and, importantly, warned that privatization should occur with strict regulation only in competitive markets. But in practice, African governments seeking immediate debt relief were often under significant pressure to enact quickly the policy measures set by international financial institutions. As a result, African governments often lacked the ability to regulate the pace of policy adoption, with sometimes detrimental consequences for their populations in the initial reform period.

Third, one ironic but true point is that for market-oriented reforms to be effective, their implementation requires stable and committed governments with a high level of social and political capital. The reforms often placed an overwhelming emphasis on macroeconomic stability and market-oriented changes without adequate provision of social safety nets that contributed to weaken governments and undermine the reform agenda. The reforms of the 1980s and 1990s were often viewed as an infringement on the national sovereignty of countries, which spurred deep resentment among many governments and populations. Policy adoption itself is inevitably a political affair, a seemingly obvious fact that has largely been ignored in previous analyses of Washington Consensus policy reforms (Mkandawire and Olukoshi 1995; Mkandawire and Soludo 1999; Mkandawire 1999; Herbst 1990). While international financial institutions often attributed the lack of success with the reform agenda to weak state capacity, the focus on market orientation and limiting state intervention in development activities led to market failures. State intervention was actually important to implement successful market-oriented reforms in some cases (Mkandawire 1999).

Fourth, it is not obvious that the market-oriented reforms emphasized by international financial institutions are the best or only route to successful economic development. Skeptics of market-oriented reforms in African point out that in many successful development efforts around the world, including many countries across Asia, governments played a prominent role for much of the critical phase of their economic development. Historically, many of today's developed economies did not fully embrace free market economies in the earlier phases of their economic development, which instead involved substantial state involvement including industrial subsidies and infant industry protection (for a discussion of the development experience of today's advanced economies, one useful starting point is Chang 2002). In Africa, many of these same practices used at other places and times were frowned upon by proponents of market-oriented policies. But before countries of sub-Saharan Africa fell into the debt crisis of the 1980s, many of them had experienced success in the period immediately post-independence in the 1960s and 1970s (Mkandawire 1999). Indeed, some of the policies that were abandoned in favor of market-oriented reforms had rational, development-motivated justifications. For

example, African states promoting low interest rates sought to boost investment and capital accumulation, and market-oriented reform of financial systems with limited competition hindered this objective. Many countries offered subsidies to the agricultural sector, although inefficient, that kept prices low to facilitate access to food for many who lived in poverty and to reduce the risk of social unrest. Protests against food prices erupted following the removal of subsidies in the 1980s and 1990s.

As general guide moving forward, we offer a few lessons from Africa's experience with the Washington Consensus reforms. First, while market-oriented reforms can be beneficial for growth, each reform policy needs to be carefully considered against institutional contexts, initial conditions of development, and socio-political environments, among other factors. Second, ownership of the reform agenda by local government with stakeholder buy-in is important to encourage support for the reforms and to increase the likelihood of success. Third, the negative spillovers of reform policies need to be minimized: for example, investment in social safety nets is a crucial part of reforms to protect the most vulnerable populations within the countries. Fourth, where reforms aim to achieve macroeconomic stability, they should not trade away social investment in human capital like education and health. Finally, reforms should be a process of continuous reevaluation, adjustment, and recalibration over the reform period. A reform agenda must be approached with flexibility.

■ *The authors gratefully acknowledge helpful input from Peter Henry and comments from Anusha Chari as well as outstanding research support from Christopher Heitzig and Gloria Kebirungi.*

References

Abegaz, Berhanu. 1999. "Aid and Reform in Ethiopia." Aid and Reform in Africa Project of the World Bank 35725.

Adedeji, Adebayo. 1999. "Structural Adjustment Policies in Africa." *International Social Science Journal* 51 (162): 521–28.

Archibong, Belinda. 2018. "Historical Origins of Persistent Inequality in Nigeria." *Oxford Development Studies* 46 (3): 325–47.

Ariyo, Ademola, and Afeikhena Jerome. 1999. "Privatization in Africa: An Appraisal." *World Development* 27 (1): 201–13.

Bayliss, Kate, and Christopher Cramer. 2003. "Privatisation and the Post-Washington Consensus." In *Development Policy in the Twenty-First Century: Beyond the Post-Washington Consensus*, edited by Ben Fine, Costas Lapavitsas, and Jonathan Pincus.

Benin, Samuel, Ephraim Nkonya, Geresom Okecho, John Pender, Silim Nahdy, Samuel Mugarura. Edward Kato. 2007. "Assessing the Impact of the National Agricultural Advisory Services (NAADS)

in the Uganda Rural Livelihoods. International Food Policy Research Institute Discussion Paper 00724.

Birdsall, Nancy, Felipe Valencia Caicedo, and Augusto de la Torre. 2010. "The Washington Consensus: Assessing a Damaged Brand." Center for Global Development Working Paper 213.

Chang, Ha-Joon. 2002. "Breaking the Mould: An Institutionalist Political Economy Alternative to the Neo-Liberal Theory of the Market and the State." *Cambridge Journal of Economics* 26 (5): 539–59.

Chari, Anusha, Peter Blair Henry, and Hector Reyes. 2020. "The Baker Hypothesis."NBER Working Paper 27708.

Coulibaly, Brahima S. 2017. "In Defense of the 'Africa Rising' Narrative." *Africa in Focus*, June 27. https://www. brookings. edu/blog/africa-in-focus/2017/06/27/in-defense-of-the-africa-rising-narrative/.

Coulibaly, Brahima S., Dhruv Gandhi, and Lemma W. Senbet. 2019. "Is Sub-Saharan Africa Facing Another Systemic Sovereign Debt Crisis?" Africa Growth Initiatives at Brookings Policy Brief.

Deaton, Angus. 1999. "Commodity Prices and Growth in Africa." *Journal of Economic Perspectives* 13 (3): 23–40.

Devarajan, Shantayan, David R. Dollar, and Torgny Holmgren. 2002. *Aid and Reform in Africa: Lessons from Ten Case Studies*. Washington, DC: The World Bank.

Dijkstra, A Geske, and Jan Kees Van Donge. 2001. "What Does the 'Show Case' Show? Evidence of and Lessons from Adjustment in Uganda." *World development* 29 (5): 841–63.

Djimeu, Eric W. 2018. "The Impact of the Heavily Indebted Poor Countries Initiative on Growth and Investment in Africa." *World Development* 104: 108–27.

Dollar, David, and Jakob Svensson. 2000. "What Explains the Success or Failure of Structural Adjustment Programmes?" *The Economic Journal* 110 (466): 894–917.

Due, Jean M., and Christina H. Gladwin. 1991. "Impacts of Structural Adjustment Programs on African Women Farmers and Female-Headed Households." *American Journal of Agricultural Economics* 73 (5): 1431–39.

Easterly, William. 2000. "The Effect of IMF and World Bank Programs on Poverty." Unpublished.

Easterly, William. 2019. In Search of Reforms for Growth: New Stylized Facts on Policy and Growth Outcomes. NBER Working Paper 26318.

Ekpo, Akpan H. 1992. "Economic Development under Structural Adjustment: Evidence from Selected West African Countries." *Journal of Social Development in Africa* 7 (1): 25–43.

Elson, Diane. 1995. "Gender Awareness in Modeling Structural Adjustment." *World Development* 23 (11): 1851–68.

Fernández, Andrés, Stephanie Schmitt-Grohé, and Martín Uribe. 2020. "Does the Commodity Super Cycle Matter?" NBER Working Paper 27589.

Franses, Philip Hans, and Eva Janssens. 2018. "Inflation in Africa, 1960–2015." *Journal of International Financial Markets, Institutions and Money* 57: 261–92.

Grier, Kevin B., and Robin M. Grier. 2020. "The Washington Consensus Works: Causal Effects of Reform, 1970–2015." *Journal of Comparative Economics* 49 (1): 59–72.

Herbst, Jeffrey. 1990. "The Structural Adjustment of Politics in Africa." *World Development* 18 (7): 949–58.

Hickey, Sam. 2013. "Beyond the Poverty Agenda? Insights from the New Politics of Development in Uganda." *World Development* 43: 194–206.

Ihonvbere, Julius O. 1993. "Economic Crisis, Structural Adjustment and Social Crisis in Nigeria." *World Development* 21 (1): 141–53.

Lynch, Gabrielle, and Gordon Crawford. 2011. "Democratization in Africa 1990–2010: An Assessment." *Democratization* 18 (2): 275–310.

Mkandawire, Thandika. 1999. "The Political Economy of Financial Reform in Africa." *Journal of International Development* 11 (3): 321–42.

Mkandawire, P. Thandika, and Charles Chukwuma Soludo. 1999. *Our Continent, Our Future: African Perspectives on Structural Adjustment*. Asmara, Eritrea: Africa World Press, Inc.

Mkandawire, Thandika, and Adebayo Olukoshi. 1995. *Between Liberalisation and Oppression: The Politics of Structural Adjustment in Africa*. Dakar, Senegal: CODESRIA.

Naiman, Robert, and Neil Watkins. 1999. *A Survey of the Impacts of IMF Structural Adjustment in Africa: Growth, Social Spending, and Debt Relief*. Preamble Center.

Okonjo-Iweala, Ngozi. 2014. *Reforming the Unreformable: Lessons from Nigeria*. Cambridge, MA: MIT Press.

Onyekwena, Chukwuka, and Mma Amara Ekeruche. 2019. "Is a Debt Crisis Looming in Africa?" *Africa in Focus*, April 10. https://www.brookings.edu/blog/africa-in-focus/2019/04/10/is-a-debt-crisis-looming-in-africa/.

Oqubay, Arkebe. 2018. "Industrial Policy and Late Industrialization in Ethiopia." African Development Bank Group Working Paper 303.

Parker, David, and Colin Kirkpatrick. 2005. "Privatisation in Developing Countries: A Review of the Evidence and the Policy Lessons." *Journal of Development Studie*s 41 (4): 513–41.

Prati, Alessandro, Massimiliano Gaetano Onorato, and Chris Papageorgiou. 2013. "Which Reforms Work and under What Institutional Environment? Evidence from a New Data Set on Structural Reforms." *Review of Economics and Statistics* 95 (3): 946–68.

Rodrik, Dani. 2006. "Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's Economic Growth in the 1990s: Learning from a Decade of Reform." *Journal of Economic Literature* 44 (4): 973–87.

Rwamigisa, Patience B., Regina Birner, Margaret N. Mangheni, and Arseni Semana. 2018. "How to Promote Institutional Reforms in the Agricultural Sector? A Case Study of Uganda's National Agricultural Advisory Services (NAADS)." *Development Policy Review* 36 (5): 607–27.

Sachs, Jeffrey D., and Andrew M. Warner. 1995. "Economic Convergence and Economic Policies." NBER Working Paper 5039.

Sender, John. 1999. "Africa's Economic Performance: Limitations of the Current Consensus." *Journal of Economic Perspectives* 13 (3): 89–114.

Serra, Narcís, and Joseph E. Stiglitz. 2008. *The Washington Consensus Reconsidered: Towards a New Global Governance.* Oxford: Oxford University Press.

Stiglitz, Joseph E. 2005. "More Instruments and Broader Goals: Moving toward the Post-Washington Consensus." In *Wider Perspectives on Global Development*, edited by UNU-Wider, Anthony B. Atkinson, Kaushik Basu, Jagdish N. Bhagwati, Douglass C. North, Dani Rodrik, Frances Stewart, Joseph E. Stiglitz, and Jeffrey G. Williamson, 16–48. London: Palgrave Macmillan.

Tekeste, Abraham. 2014. *Trade Policy and Performance of Manufacturing Firms in Ethiopia.* Scholars' Press.

Williamson, John. 1993. "Democracy and the 'Washington Consensus'." *World Development* 21 (8): 1329–36.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 157–174

Statistical Significance, *p*-Values, and the Reporting of Uncertainty

Guido W. Imbens

Theodor Geisel, better known by the nom de plume Dr. Seuss, published *The Sneetches* in 1961. In this children's story, the Star-Belly Sneetches viewed themselves as superior to the Plain-Belly Sneetches. When the character of Sylvester McMonkey McBean arrives with a machine that can add or remove belly stars (for a modest fee), social upheaval results. In empirical work in economics, stars have long been attached to numbers in tables and figures to indicate the level of statistical significance: one star typically refers to an estimate that is statistically significant at a 10 percent level; two stars, the 5 percent level; and the coveted three stars, the 1 percent level. In the word of Dr. Seuss: "Those stars weren't so big. They were really so small./You might think such a thing wouldn't matter at all./But, because they had stars, all the Star-Belly Sneetches/Would brag, 'We're the best kind of Sneetch on the beaches.'" In empirical studies, estimates with one, two, or three stars are often viewed as superior to those without such adornments.

The statistical significance indicated by stars in tables of empirical results is a concept that is at the same time widely used, widely misunderstood, and widely decried, probably more than any other statistical notion. In this essay, I begin with a short overview of the current controversies among some academic journals and professional societies in reporting *p*-values and statistical significance. Some

■ *Guido W. Imbens is Professor of Economics, Graduate School of Business, Professor of Economics, Department of Economics, and Senior Fellow, Stanford Institute for Economic Policy Research (SIEPR), all at Stanford University, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is imbens@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.157.

journals, following the "star-off" machine of Sylvester McMonkey McBean, have started removing indicators of statistical significance before publication, or even further, any use of hypothesis testing.

I then turn to three distinct concerns that have been raised—against or going even further to disallow—the use of statistical significance and *p*-values. The first concern is that often *p*-values and statistical significance do not answer the question of interest. In many cases, researchers are interested in a point estimate and the degree of uncertainty associated with that point estimate as the precursor to making a decision or recommendation to implement a new policy. In such cases, the absence or presence of statistical significance (in the sense of being able to reject the null hypothesis of zero effect at conventional levels) is not relevant, and the all-too-common singular focus on that indicator is inappropriate. Statistical education has arguably failed in clarifying to decision makers, even those with a reasonable degree of statistical sophistication, the key issues involved in decision making under uncertainty.

The second concern arises if a researcher is legitimately interested in assessing a null hypothesis versus an alternative hypothesis: say, the efficient market hypothesis or the permanent income hypothesis. Such cases do commonly arise in economics, although perhaps not as often as in physical sciences, and certainly not as often as the prevalence of null hypothesis testing in empirical work would suggest. As Abadie (2020) writes, "in economics . . . there are rarely reasons to put substantial prior probability on a point null." Questions have been raised whether *p*-values and statistical significance are useful measures for making the comparison between the null and alternative hypotheses. The use of a uniform standard (the ubiquitous 5 percent level for statistical significance) irrespective of context has been questioned. In addition, alternatives to *p*-values have been proposed for this setting, including Bayes factors. Here, I do think there is a limited but important role for *p*-values. Although I agree with much of the sentiment that small *p*-values are not *sufficient* for concluding that the null hypothesis should be abandoned in favor of the alternative hypothesis, I do think that small *p*-values are *necessary* for such a conclusion. More specifically, in cases where researchers test null hypotheses on which we place substantial prior probability, it is difficult to see how one could induce anyone to abandon that belief without having a very small *p*-value. Reporting such a *p*-value would seem a reasonable way to summarize evidence.

The third concern is the abuse of *p*-values. Because in practice much importance is attached to small *p*-values and statistical significance—the number of stars in a table—there are strong incentives for researchers to obtain more favorable *p*-values. To put it bluntly, researchers are incentivized to find *p*-values below 0.05. This has led to concerns about researchers searching for specifications (whether consciously or unconsciously) that lead to such *p*-values in ways that invalidate the meaning and interpretation of those *p*-values. This has become known as *p*-hacking. On the other side of the publication process, there are concerns that results without statistical significance are less likely to be accepted for publication. There is interesting recent work on detecting the presence of *p*-hacking and/or publication bias

(Andrews and Kasy 2019; Elliott, Kudrin, and Wuthrich 2019; Brodeur, Cook, and Heyes 2018). One approach to avoid issues of *p*-hacking relies on the use of pre-analysis plans in which a researcher specifies in advance how data will be gathered and analyzed (Casey, Glennerster, and Miguel 2012; Duflo et al. 2020; Olken 2015), as supported by the AEA registry for randomized experiments.

In this essay, I argue that I find the first concern the most compelling. Statistical significance has been over-emphasized in empirical research.[1] In many cases where decision makers are faced with deciding whether to implement a new policy or not, confidence intervals are a more useful way of communicating uncertainty of point estimates. It would be even better, in my view, to report Bayesian posterior intervals, but in many cases confidence intervals can be interpreted as posterior intervals, and so this often becomes a minor quibble. In cases where Bayesian posterior intervals and confidence intervals differ substantially, I would more strongly prefer posterior intervals.

With regard to the second issue, in which cases where the questions of interest are naturally formulated as hypothesis tests, I think that advantages of Bayes factors over *p*-values are relatively minor. In such cases, it is my view that *p*-values are a reasonable and standardized way of communicating the strength of the evidence.[2] Summarizing the strength of that evidence by a binary indicator—whether a statistically significant at the 5 or 1 percent level—seems to serve little purpose.

Concerning the third issue, *p*-hacking, it would be useful both to lower the incentives for *p*-hacking by de-emphasizing statistical significance thresholds (not reporting stars in tables), and to make it more difficult to *p*-hack by rewarding pre-analysis plans whenever feasible.

Given that this debate over statistical significance and *p*-values has gone on for a long time, I will say little that is new, and perhaps little that is controversial. My aim is to help readers understand the basic issues and why various recommendations have been made in the literature. Cox (2020) offers another recent discussion of some of these issues.

Controversy about the Reporting of *p*-values and Significance Levels

Despite the widespread use of statistical significance and *p*-values, there is much controversy in the academic literature over its appropriate role. Many authors—including multiple journal editors in empirical fields (as opposed to journals devoted to theoretical statistics)—have weighed in on the merits of reporting (in decreasing order of controversy) statistical significance, *p*-values, confidence

[1] The alleged importance of statistical significance has even entered into fiction, as in Nesbø's (2012, p. 93) crime novel *The Bat*: "Trying to find a pattern . . . is hopeless without statistics. Cold, concise statistics. Keyword number one is statistical significance. In other words, we're looking for a system that cannot be explained by statistical chance."

[2] In fact, I have written papers focused primarily on the calculation of *p*-values: including, for example, Athey, Eckles, and Imbens (2018).

intervals, and Bayesian intervals.[3] At the same time, theoretical work on properties of tests continues to attract much attention. The 1995 paper by Benjamini and Hochberg on controlling the false discovery rate when multiple statistical tests are being carried out has been cited well over 70,000 times in 25 years.

The editor of the journal *Basic and Applied Social Psychology* (BASP) went the furthest in terms of restricting the reporting of tests, ultimately banning the use of significance levels, including *p*-values as well as confidence intervals. In 2014, the editor of BASP wrote, "prior to publication, authors will have to remove all vestiges of the NHSTP [Null Hypothesis Statistical Testing Procedures] (*p*-values, *t*-values, *F*-values, statements about 'significant' differences or lack thereof, and so on)" (Trafimov 2014, p. 1). The next year the editors went further and also banned confidence intervals, although, "Bayesian procedures are neither required nor banned" (Trafimow and Marks 2015, p. 1). Back in 1986, the *American Journal of Public Health* included an "Editor's Note" (1986, p. 587, in response to Fleiss 1986) that drew a line between *p*-values and confidence intervals: "We . . . have encouraged the use of confidence intervals. We believe that the quantitative message that they convey is less subject to misinterpretation than significance testing or *p*-values." Editors of some economics journals have drawn the line between reporting indicators of statistical significance and *p*-values. Both *Econometrica* and the *American Economic Review* have policies on their website discouraging the use of stars to indicate statistical significance. *Econometrica* does explicitly encourage standard errors and confidence intervals: "Please do not use asterisks or bold face to denote statistical significance. We encourage authors to report standard errors and coverage sets or confidence intervals."[4]

The actual act of banning a probability calculation in a scientific journal is quite striking. As Hal Stern (2016 p. 23) writes,

> The *p*-value is a probability calculation giving the probability of an event (observing a more extreme t statistic) under specific assumptions: The statistical model is correct and H_0 is true. Probability calculations do not seem particularly objectionable. Why then would BASP [*Basic and Applied Social*

[3]To know the views of the authors, it often suffices to read the titles of such editorials or articles. A partial list of examples includes: "P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin," in *Journal of Clinical Epidemiology* (Feinstein 1998); "The Cult of Statistical Significance" (Ziliak and McCloskey 2011); "That Confounded P-value," in *Epidemiology* (Lang, Rothman, and Cann 1998); "A Dirty Dozen: Twelve *P*-value Misconceptions" (Goodman 2008); "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values" (Colquhoun 2014); "Toward Evidence-Based Medical Statistics. 1: The *P* value Fallacy" (Goodman 1999a); "The End of the p value" (Evans, Mills, and Dawson 1988); "The Difference between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant" (Gelman and Stern 2006); "Confidence Intervals Rather than P values: Estimation Rather than Hypothesis Testing" (Gardner and Altman 1986); "In Praise of Confidence Intervals" (Romer 2020); and "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (Berger and Sellke 1987). In "Why Most Published Research Findings Are False," Ioannidis (2005) writes: "Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values."

[4]This policy predates my term as Editor of *Econometrica*, and I had no involvement in its formulation.

> *Psychology*] ban *p*-values? . . . It is true that *p*-values are often misinterpreted and abused . . . but that by itself does not seem like a compelling reason to ban them.

Perhaps even more striking, the American Statistical Association put out an official statement on *p*-values that included the following (Wasserstein and Lazar 2016):

> Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.[5]

It is surely quite unusual for a professional society to weigh in on a specific scientific issue like the merit of a given statistic. In a blog post on the website of *Nature,* Monya Baker (2016, p. 151) writes: "'This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics,' says executive director Ron Wasserstein. 'The society's members had become increasingly concerned that the *p*-value was being misapplied in ways that cast doubt on statistics generally,' he adds."

A subsequent article by Wasserstein, Schirm, and Lazar (2019, p. 1), although not a formal statement of the American Statistical Association, went further than the original words of caution by explicitly recommending against the use of statistical significance indicators:

> The ASA Statement on *p*-values and statistical significance stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

To put this in perspective, I find it difficult to imagine the American Economic Association issuing an edict that a certain statistical approach would be banned (say, the use of instrumental variables) or the editor of the *American Economic Review* prohibiting researchers from mentioning a method or economic theory (say, the

[5] In the spirit of full disclosure, I was part of the committee that was tasked with crafting the statement.

permanent income hypothesis or rational expectations) solely because the editor felt that these methods or theories have at times been misapplied.

Although the use of statistical significance is common in economics, these discussions about statistical significance and *p*-values have not generated quite as much excitement in the economics profession as in other fields using statistical methods. In one recent exception, David Romer (2020) carefully documents that the majority of empirical papers in three leading economics journals (*American Economic Review*, *Quarterly Journal of Economics*, and *Journal of Political Economy*) focus primarily on point estimates and statistical significance in various forms. He argues against this practice and recommends reporting confidence intervals instead to summarize the uncertainty in the point estimates: "Focusing on point estimates and statistical significance obscures the implications of the findings for those values [values other than the point estimate and zero]. In addition, as discussed below, this focus also leaves out important information even about the strength of the evidence against a parameter value of zero." Another exception in the economics literature is Abadie (2020), who points out that in some cases, nonsignificant results may be much more informative than significant results in terms of changing beliefs about plausible values of the parameters of interest.

Estimation versus Hypothesis Testing

I will begin with some comments about the general nature of empirical work in economics and the relative importance of estimation versus hypothesis testing. Although hypothesis testing is routinely used in economics, I would submit that many of the substantive questions are primarily about point estimation and their uncertainty, rather than about testing. However, many studies where estimation questions should be the primary focus present the results in the form of hypothesis tests. Romer (2020) presents a specific example—the return to schooling—where testing a null hypothesis of no effect is common, yet arguably of little or no substantive interest. One would be hard-pressed to find an economist who believes that the return to education is zero. As Romer (p. 56) notes, "[T]he vast previous work in this area already provides overwhelming evidence that the rate of return is positive." Imagine for a moment that the abstract of a paper in an economics journal claimed, along the lines of the abstracts of many medical papers: "We show that an increase in education causes significantly higher earnings." One rarely sees such abstracts, because such a finding would not be surprising or interesting. For the same reason, such claims should not feature prominently in the paper. What is of interest in such papers is the magnitude and uncertainty of the estimates, and the robustness to identification concerns, not whether the data allow for the rejection of a zero effect.

Given this distinction between estimation and testing problems, in the next two sections I will discuss the role of *p*-values and statistical significance in analyses for such problems.

Decision Making under Uncertainty

Consider a decision maker choosing whether to implement a new policy—perhaps mandating a new early childhood educational program (Krueger and Whitmore 2001; Schanzenbach 2006; Chetty et al. 2011), or making micro credit available to communities in developing countries (Banerjee, Karlan, and Zinman 2015; Crépon et al. 2015; Meager 2019), or changing a search algorithm for a tech company (Gomez-Uribe and Hunt 2015; Gupta et al. 2019). Suppose the only unknown component of the utility of implementing the policy is the average treatment effect (the difference in the average outcome if everybody was exposed to the intervention versus the average outcome if nobody was exposed). To inform this decision, suppose that a randomized experiment was conducted. In this experiment, a sample of units is randomly divided into two sub-samples, with units in the first sub-sample exposed to the intervention and the units in the second sub-sample exposed to the old regime. (I am focusing here on an example with a randomized experiment because it abstracts from some other concerns about internal validity that would also come up in such discussions in the absence of randomization: for general discussions of these issues, see Deaton 2010; Imbens 2010, 2018; and Deaton and Cartwright 2018.)

A question is what information should the statisticians bring to the meeting with the (sophisticated) decision maker after having analyzed the data. In my experience, it is common in such settings for the statistician to present point estimates of the average effect, together with some combination of statistical significance, standard errors, confidence intervals, subgroup analyses, and robustness checks. A discussion might then ensue concerning the magnitude of the effect and the precision of the estimated effect, where the latter discussion would cover the degree of statistical significance and standard errors. There would also be a discussion regarding the credibility of the findings (especially in settings where the estimates are not based on randomized experiments), as well as their external validity and any evidence of heterogeneity. Kohavi, Henne, and Sommerfield (2007), Kohavi, Tang, and Xu (2020), and Gupta et al. (2019) discuss in more detail the process of decision making in the context of randomized experiments in a business setting. Kohavi views experiments in this setting, and data-driven decision making more generally, as helping reduce the importance of what he has called the Highest Paid Person's Opinion (HIPPO) in less formal versions of these discussions.

In this setting of providing information to decision makers, I want to make two claims. First, what is most relevant for the decision maker is the point estimate with some measure of the uncertainty of that point estimate, and some sense of the robustness and identification issues. The second claim is that the testing of statistical hypotheses—and thus the reporting of p-values or statistical significance—is essentially irrelevant in this case. The common practice of prominently reporting these measures is therefore largely misguided. As the statement of the American Statistical Association claims, correctly in my view, "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold" (Wasserstein and Lazar 2016).

To provide further support for the view that in this case the appropriate focus is on point estimates and measures of uncertainty, consistent with the view of some econometricians of econometrics as applied decision theory (Leamer 1978; Chamberlain 2000; Manski 2013; Hirano 2010; Dehejia 2005), let me make this example more specific. Suppose the point estimate of the treatment effect is $\hat{\tau} > 0$ (with positive values preferred relative to negative values by the decision maker), and suppose the standard error is σ. Let us also suppose that the analysts are confident that the sampling distribution of the estimator is approximately normal, so that the 95 percent confidence interval is plus or minus 1.96 standard deviations from the point estimate $\hat{\tau}$. Given these numbers, the discussion of the decision makers would typically center on the plausibility of the estimates, the magnitude of the cost relative to the estimated benefits, the external validity of the estimates (will they actually generalize to the population they might be applied to), evidence of heterogeneity in the effects, and the possibility (or explicitly, the probability) of effect sizes that would render the decision to be clearly wrong after it was taken, possibly taking into account prior beliefs. These topics have an implicitly Bayesian flavor: the decision maker is in various ways confronting the point estimates with prior beliefs. The use of confidence intervals as the basis for a discussion in a Bayesian spirit is (approximately) justified by the interpretation of the confidence intervals as Bayesian intervals, although this is rarely made explicit.[6]

In addition, identication issues may arise, for example, from lack of randomization, or via uncertainty about differences between the study population and the target population, or uncertainty about differences between the future and the past. These are often dealt with informally by just acknowledging that some degree of additional uncertainty exists, rather than by using more principled ways of calculating bounds along the lines of the work by Manski (2013).

Although the topic of statistical significance is often brought up in these discussions, it often is used inappropriately by implicitly interpreting insignificant estimates as true zeros. To illustrate the lack of a role for the significance level, suppose the utility from the general implementation of the treatment is equal to the true treatment effect, so that implicitly the cost of implementing the treatment is zero, and there is no risk aversion. In this case, the right decision given a treatment effect equal to τ would be to implement the intervention if the estimated value of $\tau > 0$, and not otherwise. From a Bayesian perspective, the only reason not to implement the intervention given a positive estimate $\hat{\tau}$ would be that the prior distribution for τ implies that the posterior expected value for τ is

[6]This is based on the Bernstein-Von Mises theorem that, informally, says that in many cases confidence intervals can be viewed as approximate Bayesian posterior intervals (Van der Vaart 2000). Although there are multiple settings where confidence intervals are not based on asymptotic normality (for example, in instrumental variables settings with weak instruments, or with settings with unit roots), I have not seen analysts attempt to explain such confidence intervals to policy-makers, and I would expect that to be a challenging task. In such cases where the Bernstein-von Mises Theorem does not hold and confidence intervals are *not* similar to (Bayesian) posterior intervals I would strongly prefer the Bayesian intervals over confidence intervals. See Sims and Uhlig (1991) for a related discussion in the context of unit roots.

negative, despite that positive value for the estimated $\hat{\tau}$. If one really believes that a flat prior is appropriate, then even the value of the standard error σ does not actually matter. In practice, of course, a flat prior is almost always implausible and the prior standard deviation is often modest. Moreover, one may a priori be skeptical about the proposed intervention, so that the prior mean is negative. In that case, one needs not just a positive point estimate, but also a sufficiently positive and precise point estimate to justify the implementation of the proposed intervention. In some cases, such a prior distribution could be justified more systematically using data from prior experiments using an empirical Bayes approach (Morris 1983). Although I am pushing for a more Bayesian approach than is typically reported, I would be comfortable with the statisticians just reporting the point estimates and confidence intervals, because decision makers can then combine that with their own prior distributions (for example Andrews and Shapiro 2020).

In the case I just outlined, presenting the implicitly Bayesian decision makers with *p*-values or conventional indicators of statistical significance does them a disservice and in practice underestimates their sophistication. In practice it often leads decision makers to act as if statistically insignificant results are truly zero. In doing so, it confuses the matter at hand by distracting the decision maker from the real issues: what are the costs of type I and type II errors, what are their prior beliefs, and how much the estimates change those beliefs. As Abadie (2020) shows, statistical significance need not change those beliefs very much.

Assessing the Relative Merits of the Null Hypothesis versus an Alternative Hypothesis

Although I have argued that in many cases point estimates and confidence intervals are the most useful summary statistics from a statistical analysis, there are settings in economics where it may be reasonable to focus on testing null hypotheses, often about a particular economic theory. We may be interested in testing the permanent income hypothesis, the efficient market hypothesis, whether there are constant returns to scale, whether there is a "sheepskin effect" of graduation in the returns to education, or whether particular groups are discriminated against. Although in all these examples one can still argue how seriously to take such a sharp null hypothesis (that is, with sufficiently large samples we might expect to reject most of such hypotheses), it may still be useful to assess whether there is clear evidence in the available data against such theories. To make the discussion specific, let me focus on an (non-economics) example where testing whether the null hypothesis holds may be more relevant than the magnitude of deviations from the null hypothesis if it is violated, and where the testing has generated much controversy. A similar example is the hot-hand fallacy (Ritzwoller and Romano 2020).

This example attracted great controversy in the psychology literature. In the *Journal of Personality and Social Psychology*, Bem (2011) studies whether precognition

exists: that is, whether future events retroactively affect people's responses. Reviewing nine experiments, he finds (from the abstract): "The mean effect size (d) in psi performance across all nine experiments was 0.22, and all but one of the experiments yielded statistically significant results." This finding sparked considerable controversy, some of it methodological. The title of a response by Wagenmakers et al. (2011) sums up part of the critique: "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." A *New York Times* article on the controversy was titled, "Journal's Paper on ESP Expected to Prompt Outrage," which states: "Many statisticians say that conventional social-science techniques for analyzing data make an assumption that is disingenuous and ultimately self-deceiving: that researchers know nothing about the probability of the so-called null hypothesis" (Carey 2011). The same issue is addressed in the statement by the American Statistical Association: "By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis" (Wasserstein and Lazar 2016).

In this case, it would appear there is reasonable interest in testing the sharp null hypothesis irrespective of the magnitude of the effect: that is, the question of whether precognition exists at all is interesting. The same can be argued for drug trials, where some cases have found that a particular drug or medical procedure has some effect on a medical condition, even if the effect is very small and possibly far below a cost-effective level. Such a finding is informative about possible mechanisms and may suggest further research into alternative treatments. I see these settings as qualitatively different from the decision problem discussed in the previous section, where the question was whether to implement a particular intervention. Here the decision question is whether to investigate a particular scientific question further. In this setting I disagree with Neyman's (1935) comment that knowing that a treatment has some effect, even if the average effect is zero, is purely academic. Here, such a finding is important even if it is not of immediate policy relevance.

Even if we agree that assessing the null hypothesis relative to an alternative hypothesis is for certain questions a matter of interest, one might argue as to whether the *p*-value is the most useful statistic for assessing that null hypothesis. Arguments have been put forward in favor of an explicitly Bayesian approach, as in, for example, Wagenmakers et al. (2011), Goodman (1999b), and Carey (2011). Using a probability that a null hypothesis that precognition does not exist equal to 10^{-20} (a prior distribution more or less in agreement with my own), Wagenmakers et al. (2011) show that the posterior probability that precognition exists, given some of Bem's experiments, remains very small so as to make it unlikely. I agree with the premise of Wagenmaker's argument that a small *p*-value alone is not *sufficient* to reject the null hypothesis in favor of the alternative hypothesis. However, I do think a small *p*-value is *necessary* for this. It is difficult to imagine a dataset that would contain enough information to reject the null hypothesis of no precognition *without* a small *p*-value. Here I agree with Benjamini (2016, p. 1) who writes: "[The *p*-value] offers a first line of defense against being fooled by randomness, separating signal from noise."

There is a substantial literature on whether the use of a "Bayes factor" would be more informative than p-values, part of an even larger literature on alternatives to p-values.[7] Given a null hypothesis and an alternative hypothesis, the Bayes factor is the ratio of the marginal likelihood of the data under the null hypothesis and the marginal likelihood of the data under the alternative hypothesis (Kass and Raftery 1995). Unlike the fully Bayesian calculation of the posterior probability that the null hypothesis is true given the data, the calculation of a Bayes factor does not require a prior probability that the null hypothesis is true. A couple of points are worth noting about this measure of the evidence. First, an attractive feature of the Bayes factor is that it is symmetric in its dependence on the two hypotheses, whereas the p-value conditions on the null hypothesis being true. Second, to calculate the actual probability of one of the hypotheses being true, the Bayes factor is not sufficient: we also need the prior probabilities of either hypotheses being true. Such prior probabilities are likely to be controversial. Finally, and this is probability the biggest reason the use of the Bayes factor is less common in practice than the p-value, it also requires a prior distribution to deal with nuisance parameters. For example, if the null hypothesis is sharp—say, that a coin is fairly balanced between heads and tails—the alternative hypothesis is typically not sharp: all values for p other than $p = 1/2$ are consistent with the alternative hypothesis. The calculation of the Bayes factor requires the specification of a prior distribution under the alternative hypothesis, that is, a prior distribution for p on the interval [0,1] excluding the value 1/2. Although in specific cases there may be natural prior distributions to consider (for some discussions, see Goodman 1999b; Berger and Pericchi 1996), in general this makes the Bayes factor calculations more challenging and controversial. For example, if we wish to test the null hypothesis that a drug has no effect on a health outcome, there is no natural prior distribution for the treatment effect under the alternative hypothesis. In the end, I do not see the advantages of Bayes factors over p-values as sufficient to convince researchers to adopt this technology more widely.

Finally, if one is comfortable with the use of p-values in settings such as these, the question remains whether the use of a standardized threshold of 5 percent is useful to indicate statistical significance. At some level, it is not surprising that researchers adopt a standard—whether 5 percent or some other level—to facilitate communication. However, it is difficult to justify a single standard across a wide range of applications that may differ enormously: for example, in terms of size of datasets, costs of type I and type II errors, the number of tests performed, and the prior beliefs about the null hypotheses. Such concerns have led researchers in genetics to move to substantially lower significance thresholds (Storey and Tibshirani 2003). In high-energy physics, statistical significance is commonly ascribed only

[7] As one example, "Lindley's paradox" concerns the discrepancy between frequentist testing and Bayesian calculations of the probability that the null hypothesis is true. The paradox is that for a given significance level p, a test can be statistically significant, even though the posterior probability of the null hypothesis can be arbitrarily high. This can happen when the prior probability of the null hypothesis is non-negligible, the sample is large, and the prior distribution over values consistent with the alternative hypothesis is sufficiently spread out.

to findings with *p*-values below 3×10^{-7}, corresponding to estimates more than five standard errors away from zero (for example, Sinervo 2002). Benjamin et al. (2018) suggest using 0.005 (corresponding approximately to estimates more than three standard errors away from zero), rather than 0.05, as a standard for indicating statistical significance in cases where the question of interest is whether to override a strong prior belief.[8]

Publication Bias and *p*-hacking

For academic researchers, the presence or absence of a statistically significant result may influence the chance of publication and thus career success. For drug companies, a *p*-value less than or more than 0.05 can mean a difference in revenues of billions of dollars. Thus, researchers may be tempted to shape or change their analyses to reach the unstated goal of a statistically significant result.

One of the most striking examples of such abuse is that of Scott Harkonen, the fomer CEO of InterMune. Intermune did a randomized trial for a drug that Harkonen called "a $2 billion market opportunity for InterMune" (Brown 2013). Comparing survival rates for all treated and control patients in the study led to a *p*-value of 0.08, not statistically significant at conventional levels. However, by creatively looking for subgroups (who had not been included in any pre-analysis plan), InterMune found that for the subsample of participants with mild to moderate (but not severe) cases of the disease, the drug had an effect on survival with a highly significant *p*-value of 0.004. The company sent out a press release: "InterMune Announces Phase III Data *Demonstrating* [my italics] Survival Benefit of Actimmune in IPF . . . . Reduces Mortality by 70 percent in Patients with Mild to Moderate Disease." As Mayo (2020) describes this episode, which ultimately led to a conviction for issuing a misleading press report, Harkonen "reported statistically significant drug benefits had been shown, without mentioning this referred only to a subgroup he identified from ransacking the unblinded data." Indeed, Brown (2013) reports on a follow-up study carried out by InterMume that "enrolled only people with mild to moderate lung damage, the subgroup whose success was touted in the press release. And it failed. A little more than a year into the study, more people on the drug had died (15 percent) than people on placebo (13 percent). That was the death knell for the drug. Most insurers stopped paying for it."

The suspicion is that there are many more cases that do not have billions of dollars at stake, but where researchers also search for specifications that lead to *p*-values that cross the threshold into the territory that allows them to be referred to as statistically significant (Head et al. 2015). Concerns about searching through specifications for statistically significant results have been prominent in econometrics at least since the work of Edward Leamer (1978, 1983). In particular, there

[8] I am sympathetic to this proposal, and in fact was one of the many authors on this paper.

may be substantial incentives for researchers to come up with surprising findings of effects where prior beliefs put a high probability on these effects being absent. Such findings are more likely to be picked up by the popular press and, in general, gather attention as well as lead to publications in academic journals. Andrew Gelman has eloquently criticized many examples on his blog *Statistical Modeling, Causal Inference, and Social Science*, focusing on the concerns that even if researchers do not deliberately set out to calculate misleading *p*-values, they make many specification choices (the "garden of forking paths") that affect these measures, so the reported results should not be taken at face value (Gelman and Loken 2013).

One example that Gelman presents involves the "hurricanes versus himmicanes" controversy: is damage greater from hurricanes with female names rather than male names? The finding seems implausible on its face, given that female and male names are assigned to hurricanes on an alternating basis. However, Jung et al. (2014) apply a 5 percent significance standard and write in their abstract: "We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes." If the paper had been submitted to the *Proceedings of the National Academy of Sciences* with an abstract reading "We use more than six decades of death rates from US hurricanes to show that the damage of hurricanes is not related to the gender of their name," would the paper have been accepted for publication? If the authors had not found a statistically significant result, would they have simply moved on to another project?

One direction that has been explored in the literature is to assess evidence for possible abuse of *p*-values by exploring specifications that are not reported, or what is typically referred to as "*p*-hacking" (Andrews and Kasy 2019; Elliott, Kudrin, and Wuthrich 2019; Brodeur, Cook, and Heyes 2018). A related issue is publication bias, where reviewers and editors may be more inclined to accept for publication papers with low *p*-values and/or statistically significant results. The presence of *p*-hacking and publication bias can be detected using data on a large number of published articles: for example, if there is a discontinuity in the distribution of *p*-values, with a larger number of *p*-values just below 0.05 relative to the number of *p*-values just above 0.05.

Detecting *p*-hacking is one thing; addressing it is a different matter (Simmons, Nelson, and Simonsohn 2013). One possible approach is to use replication studies (as in Makel, Plucker, and Hegarty 2012), which can focus on what choices were made behind the scenes in reaching the statistically significant result. Such studies do not directly prevent *p*-hacking but can show that the announced results have less support than it might seem. De-emphasizing *p*-values (and perhaps also statistical significance more broadly) may decrease the incentives for *p*-hacking, and thus lower its prevalence. In some contexts, in particular with randomized experiments, filing a pre-analysis plans that specifies how the data will be analyzed can also help to prevent *p*-hacking (Casey et al. 2012; Chang and Li 2017; Duflo et al. 2020). Such pre-analysis plans are required by the Food and Drug Administration in its drug approval process and are becoming increasingly used in social sciences. The

American Economic Association has operated a registry for randomized experiments since 2012 that provides all the essential benefits from pre-analyses plans.

Publication bias may be more difficult to deal with. In some cases, journals are willing to pre-commit to publishing studies based on pre-analysis plans, but it is difficult to imagine that practice becoming widespread. Consider an editor approached with a proposal to investigate precognition through a well-designed, large-scale trial. Given a very strong prior belief that precognition does not exist, it is difficult to see why an editor would pre-commit to publishing such a study. On the other hand, if the study was well-designed and did find a substantial and precisely estimated effect, there would be clear arguments after the work was completed to publish such a study—if only to encourage other researchers to further investigate the topic.

Conclusion

The use of *p*-values and indicators for statistical significance has become a matter of substantial controversy. Some journals have established policies banning the use of such measures. In my view, banning *p*-values is inappropriate. As I have tried to argue in this essay, I think there are many settings where the reporting of point estimates and confidence (or Bayesian) intervals is natural, but there are also other circumstances, perhaps fewer, where the calculation of *p*-values is in fact the appropriate way to answer the question of interest. Moreover, there is little evidence that a blanket ban on *p*-values improves the quality of statistical reporting. When the journal *Basic and Applied Social Psychology* banned *p*-values, the editors wrote that, "We hope and anticipate that banning the NHSTP [null hypothesis statistical testing procedures] will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking" (Trafimow and Marks 2015, p. 2). However, a study assessing statistical studies published in the journal following the *p*-value ban concludes the opposite. Quoting from the abstract: "We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered. Readers would be largely unable to recognize this because the necessary information to do so was not readily available" (Fricker Jr. et al. 2019).

Although I do not endorse a ban on the reporting of *p*-values, I do agree that over the years, and in some disciplines more than other, *p*-values and statistical significance have been overemphasized. In many cases, the *p*-value or the measure of statistical significance is not the relevant output from an analysis of a dataset. Therefore, its prominence in the abstracts of many empirical papers is misplaced. It would be preferable if reporting standards emphasized confidence intervals (as Romer 2020 suggests) or standard errors, and, even better, Bayesian posterior intervals.

■ *I am grateful for comments by Alberto Abadie and Kei Hirano and for generous support from the Office of Naval Research through ONR grant N00014-17-1-2131.*

References

Abadie, Alberto. 2020. "Statistical Nonsignificance in Empirical Economics." *American Economic Review: Insights* 2 (2): 193–208.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Andrews, Isaiah, and Jesse M. Shapiro. 2020. "A Model of Scientific Communication." NBER Working Paper 26824.

Athey, Susan, Dean Eckles, and Guido W. Imbens. 2018. "Exact p-values for Network Interference." *Journal of the American Statistical Association* 113 (521): 230–40.

Baker, Monya. 2016. "Statisticians Issue Warning over Misuse of *p*-values." *Nature News* 531 (7593): 151.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7 (1): 1–21.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407–25.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Benjamini, Yoav. 2016. "It's Not the p-values' Fault." *The American Statistician* 70 (2).

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Berger, James O., and Luis R Pericchi. 1996. "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association* 91 (433): 109–22.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of *p*-values and Evidence. *Journal of the American statistical Association* 82 (397): 112–22.

Brodeur, Abel, Nikolai Cook, and Anthony G. Heyes. 2018. "Methods Matter: P-Hacking and Causal Inference in Economics." IZA Discussion Paper 11796.

Brown, David. 2013. "The Press-Release Conviction of a Biotech CEO and Its Impact on Scientific Research." *The Washington Post*, September 23. https://www.washingtonpost.com/national/health-science/the-press-release-crime-of-a-biotech-ceo-and-its-impact-on-scientific-research/2013/09/23/9b4a1a32-007a-11e3-9a3e-916de805f65d_story.html.

Carey, Benedict. 2011. Journal's Paper on ESP Expected to Prompt Outrage. *The New York Times*, January 06.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127 (4): 1755–1812.

Chamberlain, Gary. 2000. "Econometrics and Decision Theory." *Journal of Econometrics* 95 (2): 255–83.

Chang, Andrew C., and Phillip Li. 2017. "Preanalysis Plan to Replicate Sixty Economics Research Papers that Worked Half of the Time." *American Economic Review* 107 (5): 60–64.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126 (4): 1593–1660.

Colquhoun, David. 2014. "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values." *Royal Society Open Science* 1 (3): 140216.

Cox, David R. 2020. "Statistical Significance." *Annual Review of Statistics and its Application* 7: 1–10.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–50.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

Dehejia, Rajeev H. 2005. "Program Evaluation as a Decision Problem." *Journal of Econometrics* 125 (1–2) : 141–73.

Duflo, Esther, Abhijit Banerjee, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics. NBER Working Paper 26993.

Editor's Note. 1986. *American Journal of Public Health* 76 (5): 587–88.
Elliot, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2019. "Detecting *p*-hacking. arXiv preprint arXiv:1906.06711.
Evans, S.J., Peter Mills, and Jane Dawson. 1988. "The End of the *p*-value*? British Heart Journal* 60 (3): 177–80.
Feinstein, Alvan R. 1998. "P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin." *Journal of Clinical Epidemiology* 51 (4): 355–60.
Fleiss, Joseph L. 1986. "Confidence Intervals vs Significance Tests: Quantitative Interpretation." *American Journal of Public Health* 76 (5): 587–88.
Fricker, Ronald D. Jr, Katherine Burke, Xiaoyan Han, and William H Woodall. 2019. "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology after their *p*-value Ban." *The American Statistician* 73(S1): 374–84.
Gardner, Martin J., and Douglas G. Altman. 1986. "Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing." *Br Med J (Clin Res Ed)* 292: 746–50.
Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis Was Posited ahead of Time. http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf.
Gelman, Andrew, and Hal Stern. 2004. "The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31.
Gomez-Uribe, Carlos A., and Neil Hunt. 2015. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." *ACM Transactions on Management Information Systems* 6 (4):1–19.
Goodman, Steven N. 1999a. "Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy." *Annals of Internal Medicine* 130 (12): 995–1004.
Goodman, Steven N. 1999b. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." *Annals of Internal Medicine* 130 (12): 1005–1013, 1999b.
Goodman, Steven. 2008. "A Dirty Dozen: Twelve *P*-Value Misconceptions." *Seminars in Hematology* 45 (3): 135–140.
Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin et al. 2019. "Top Challenges from the First Practical Online Controlled Experiments Summit." *ACM SIGKDD Explorations Newsletter* 21 (1): 20–35.
Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): e1002106.
Hirano, Keisuke. 2010. "Decision Theory in Econometrics." In *Microeconometrics,* edited by Steven N. Durlauf and Lawrence E. Blume, 29–35. New York: Springer.
Imbens, Guido W. 2010. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423.
Imbens, Guido. 2018. "Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright." *Social Science & Medicine (1982)* 210: 50–52.
Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8):e124.
Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe. 2014. "Female Hurricanes Are Deadlier than Male Hurricanes." *Proceedings of the National Academy of Sciences* 111 (24): 8782–87.
Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90: 773–95.
Kohavi, Ron, Randal M. Henne, and Dan Sommerfield. 2007. "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers Not to the Hippo." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 959–67.
Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge, UK: Cambridge University Press.
Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project Star." *The Economic Journal* 111 (468): 1–28.
Lang, Janet M., Kenneth J. Rothman, and Cristina I Cann. 1998. "That Confounded P-value." *Epidemiology* 9 (1): 7–8.
Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons.
Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *The American Economic Review* 73

(1): 31–43.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. "Replications in Psychology Research: How often Do They Really Occur? *Perspectives on Psychological Science* 7 (6):537–42.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions.* Cambridge, MA: Harvard University Press.

Mayo, Deborah. 2020. "P-Values on Trial: Selective Reporting of (Best Practice Guides Against) Selective Reporting." *Harvard Data Science Review* 2 (1).

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1): 57–91.

Morris, Carl N. 1983. "Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.

Nesbø, Jo. 2012. *The Bat.* London: Random House.

Neyman, Jerzey, K. Iwaszkiewicz, and St. Kolodziejczyk.1935. "Statistical Problems in Agricultural Experimentation (with discussion)." *Journal of the Royal Statistal Society* 2 (2): 107–80.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Ritzwoller, David M., and Joseph P. Romano. 2019. "Uncertainty in the Hot Hand Fallacy: Detecting Streaky Alternatives in Random Bernoulli Sequences." Department of Statistics, Stanford University, 2019.

Romer, David. 2020. "In Praise of Confidence Intervals." *AEA Papers and Proceedings* 110: 55–60.

Schanzenbach, Diane Whitmore. 2006. "What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy* 9: 205–28.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2013. "Life after p-hacking." Paper presented at Meeting of the Society for Personality and Social Psychology, New Orleans, LA, January 13.

Sims, Christopher A., and Harald Uhlig. 1991. "Understanding Unit Rooters: A Helicopter Tour." *Econometrica: Journal of the Econometric Society* 59 (6): 1591–99.

Sinervo, Pekka K. 2002. "Signal Significance in Particle Physics." arXiv preprint hep-ex/0208005

Stern, Hal S. 2016. "A Test by Any Other Name: P Values, Bayes Factors, and Statistical Inference. Multivariate Behavioral Research." 51 (1): 23–29.

Storey, John D., and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences* 10 (16): 9440–45.

Trafimow, David. 2014. "Editorial." *Basic and Applied Social Psychology* 36 (1): 1–2, 2014.

Trafimow, David, and Michael Marks. 2015. "Editorial." *Basic and Applied Social Psychology* 37 (1):1–2.

Van der Vaart, A.W. 2000. *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L.J. van der Maas. 2011. "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100 (3): 426–32.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World beyond 'p<0.05'." *The American Statistician* 73(S1): 1–19.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2011. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor: University of Michigan Press.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 175–192

Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It

Maximilian Kasy

The author Jorge Luis Borges wrote a short story in 1941 called "The Garden of Forking Paths." The plot involves (among other elements) a journey in which the road keeps forking; a novel in which when two alternative outcomes arise, both of them happen; and a labyrinth that may or may not have been built. Statisticians have used the metaphor from Borges to convey how empirical research also involves a garden of forking paths: how data is chosen and prepared for use, what variables are the focus of inquiry, what statistical methods are used, what results are emphasized in writing up the study, and what decisions are made by journal editors about publication. If the published results are the outcome of many unobserved forking paths, then conventional estimators, hypothesis tests, and confidence sets in published studies in the social and life sciences may convey a distorted impression (Ioannidis 2005; Gelman and Loken 2013). A possible response to this issue is to "tie researchers' hands," to use another metaphor. By requiring researchers to pick beforehand which of the forking paths they will take, we might be able to restore the validity and replicability of research. Put differently, with their hands tied, researchers are prevented from cherry picking.

Faced with such concerns, applied researchers in the social and life sciences—as well as policymakers—are confronted with two sets of questions that I will address in this paper. First, how can we tell to what extent selective reporting and publication is really taking place in a given literature? How much are published estimates

■ *Maximilian Kasy is Associate Professor of Economics, University of Oxford, Oxford, United Kingdom. His email address is maximilian.kasy@economics.ox.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.175.

affected as a consequence? Second, how should we reform the practice and teaching of statistics, as well as the academic publication system, to reduce these problems?

I begin by discussing several methods which have been used in the literature to provide evidence for selective reporting and publication. These methods are based on plotting the distribution of published *p*-values, regressing published estimates on reported standard errors (or their inverse), and considering the "rate of replication" in replicated experiments (that is, the share of significant findings which are also significant when replicated). While these three methods can be useful for demonstrating the existence of selective reporting and publication, they do depend on problematic assumptions, and they allow neither estimates of the magnitude nor the form of selection. Thus, I will review two alternative methods proposed by Andrews and Kasy (2019), which allow us to estimate the extent of selective reporting by researchers and selective publication by journals. One of these approaches uses systematic replication experiments and builds on the intuition that, absent selection, original and replication estimates should be distributed symmetrically. The other approach uses meta-studies and builds on the intuition that, absent selection, the distribution of estimates should be more dispersed for findings with larger standard errors. Taken together, these approaches establish that published research in many fields is highly selected.

I will next turn to the debates about how to reform the practice of statistics and the academic publication system. As a starting point, I will argue that there are different justifiable objectives for scientific studies (Frankel and Kasy forthcoming), and that we need to be explicit about our objectives in order to discuss the tradeoffs between them. Replicability and the validity of conventional statistical inference constitute one such objective. Relevance of findings might be another objective. If our goal is to inform decision-makers or to maximize social learning, there is a strong rationale to put some emphasis on publishing surprising findings. Yet another objective could be the plausibility of published findings. If there is some uncertainty about the quality of studies and we want to avoid publishing incorrect results, we might want to put some emphasis on publishing unsurprising findings.

Against the backdrop of these different objectives, I will then discuss some current reform efforts and proposals in greater detail: for example, the push to report estimates and standard errors while de-emphasizing statistical significance, as promoted by the American Economic Association policy of banning "stars" in estimation tables, and the increasingly common requirement of pre-analysis plans which involve tying the hands of researchers in how they will analyze the data, especially in experimental research. There are also new initiatives to launch journals for null results and journals for replication studies that could fulfill an important role in a functionally differentiated publication system. They could allow for the existence of a vetted public record of findings that would be an input to meta-studies, while allowing for the existence of selective outlets with a higher profile.

In conclusion, I will argue that these debates raise some fundamental questions for statistical theory. In order to discuss these issues coherently, statistical theory should seek to understand quantitative empirical research as a social process of

communication and collective learning that involves many different actors with differences in knowledge and expertise, different objectives, and constraints on their attention and time, along with a recognition that these actors engage in strategic behavior.

Is Published Research Selected?

Forms of Selection

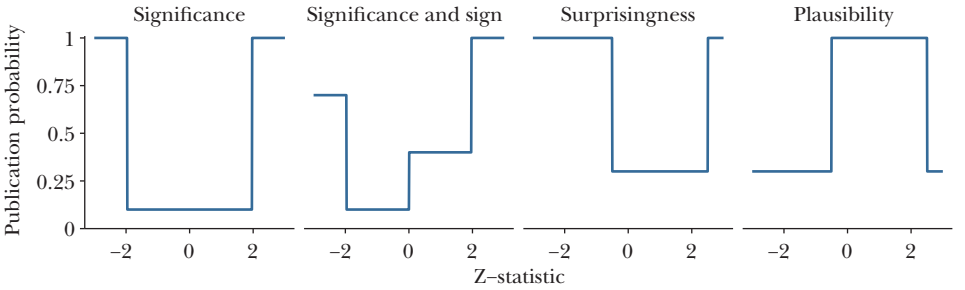
Let us begin by sketching some forms that selection based on findings might take. As noted earlier, findings might be selected by researchers as they navigate the forking paths of a research effort: which specifications are included in a paper, which outcome variables or controls are considered, and so on. Findings might also be selected by journals—for example, are null results published, or results that contradict conventional beliefs? Perhaps the most commonly discussed and criticized form of selection is based on significance. For instance, studies might be more likely to be published if their headline finding corresponds to a test-statistic exceeding the 5 percent critical value or some other conventional value.

Figure 1 illustrates different patterns of selection that might exist in the published literature. Each of the panels in this figure plots a possible dependence of the probability of publication on the z-statistic corresponding to an empirical finding, where the z-statistic is given by the estimate divided by its standard error. The relationship between the z-statistic and the probability of publication can be viewed as a reduced form summary of possible mechanisms driving selection, which might be due to various researcher or journal preferences.

For example, the left-hand panel in Figure 1 shows that if statistical significance at the 5 percent level is the key driver of what is published, then a paper is more likely to be written up if the absolute value of its z-statistic exceeds the critical value of 1.96 (for standard normal estimates): otherwise, the paper is quite unlikely to be written up and/or published. This is the pattern we found in Andrews and Kasy (2019) when analyzing data on lab experiments in economics from Camerer et al. (2016); results significant at the 5 percent level are over 30 times more likely to be published than are insignificant results in this field.

As an alternative, assume that selection occurs both on the basis of statistical significance and also based on whether an estimate has the "right sign," according to theory or conventional beliefs. In this case, as shown in the second panel, statistically significant results with the "right" sign are more likely to be published than significant results of the "wrong" sign, and in addition, statistically insignificant results with the "right" sign have some chance of being published as well. This is the pattern we found in Andrews and Kasy (2019), when analyzing data from Wolfson and Belman (2015). Studies finding a negative and statistically significant effect of minimum wage increases on employment are more likely to be published than either studies finding an insignificant effect or studies finding a positive and significant effect.

Figure 1
Some Possible Forms of Selection



Note: The first two plots show the effect of only publishing significant estimates (with a *z*-statistic above 1.96) on the bias of point estimates (average estimate minus truth) and the coverage of confidence intervals (probability of containing the truth) conditional on publication. The third plot shows the effect on the posterior absent publication.

Researchers or referees might also compare findings to a reference point other than zero. For instance, they might value surprisingness relative to some prior mean. The third panel of Figure 1 shows such a pattern in which "surprising" results are more likely to be published. As argued below, this type of pattern could be optimal when the goal of publication is to inform policy decisions. Or journal editors and referees might do the opposite, and may be disinclined to publish findings that deviate a lot from prior beliefs, because such findings are considered implausible, which might lead to selection as in the last example shown. The examples in Figure 1 are shown as step-functions for illustration only; in practice, publication probabilities might, of course, also vary continuously.

Detecting Selection

To discover the presence of selection—whether it is due to "*p*-hacking" by researchers, or due to publication bias—three methods are commonly used.

The first method is based on the *p*-values corresponding to the headline findings of a set of publications (Brodeur et al. 2016). If the distribution of these *p*-values across publications shows a discrete jump at values such as 5 percent, that provides evidence of selection. However, this method cannot spot all forms of selection, nor can it recover the form and magnitude of selection. To see why, note that the distribution of published *p*-values depends not only on selection, but also on the underlying distribution of true effects. For instance, a large number of small *p*-values, suggesting a high degree of statistical significance in the results, could be due to either a large number of null hypotheses that are indeed false, or to strong selection on the basis of significance. Observing a certain distribution of *p*-values in the published literature does not allow one to distinguish between these

two explanations. That said, without selection and for continuously distributed test-statistics such as the *t*-test, one would never expect to find a discontinuity in the density of *p*-values across studies. Such discontinuities thus do provide strong evidence of selection.

The second method for detecting selection is based on meta-studies, which regress point-estimates on standard errors (or their inverse) across a set of publications (Card and Krueger 1995; Egger et al. 1997). The meta-regression approach relies on the assumption that there is no systematic relationship between true effect size and sample size (where sample size will affect standard errors) across studies. Even under this assumption, however, many forms of selection do not create a systematic dependence between mean estimates and standard errors, and can thus not be detected in this approach. A systematic dependence between standard errors and point estimates does, however, provide evidence of selection. Additionally, meta-regressions are often used to extrapolate to the hypothetical mean estimate for a standard error of zero (corresponding to a hypothetical study with an infinite sample size). This extrapolated value is then interpreted as an estimate of the true average effect across published and unpublished studies. This interpretation is based on the implicit assumption that all studies with sufficiently large *t*-statistics are published, which implies that for small enough standard errors, all studies are published. The problem with this interpretation is that the relationship between average estimates and standard errors is never linear, but extrapolation to zero requires such a functional form restriction.

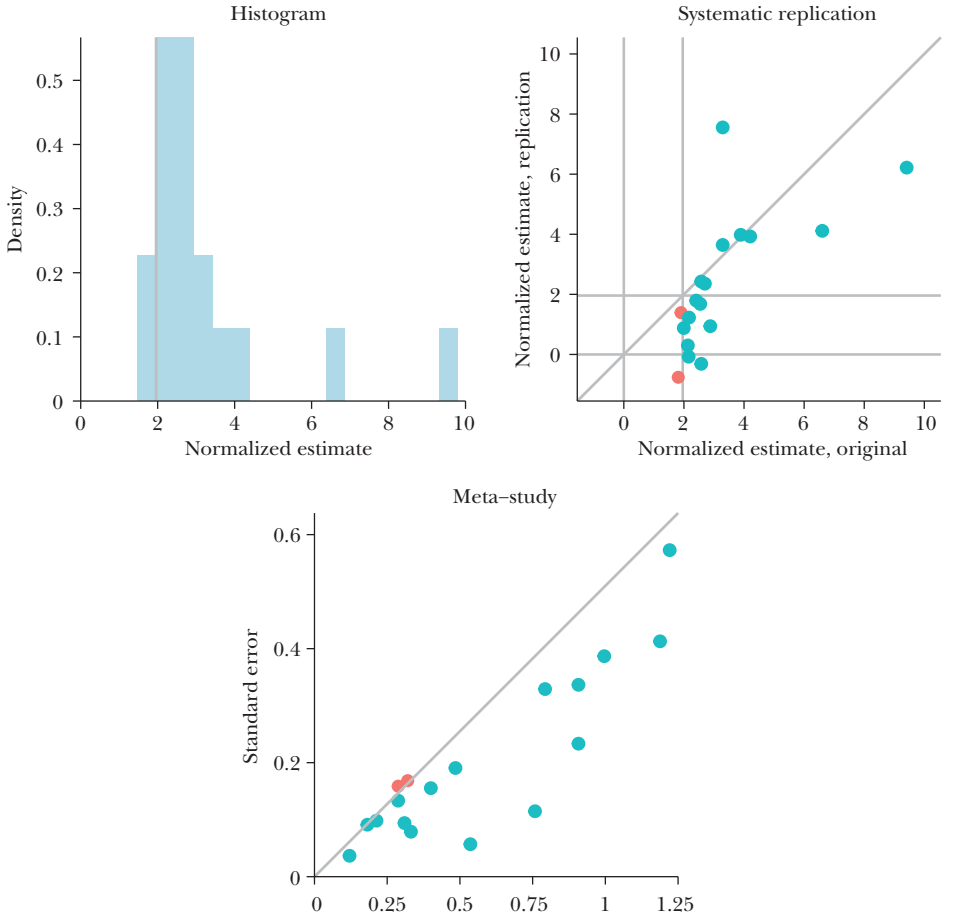
The third method of detecting selection looks at the "rate of replication" for experiments that are repeated with the same protocol, but using different subjects (Open Science Collaboration 2015). The "rate of replication" is defined as the share of published significant estimates for which the replication estimates exceed the significance threshold as well. A low rate of replication is taken as evidence of selection or some other problems. However, the "rate of replication" of significant findings, taken by itself, does not tell us much about selection. To see why, suppose first that all true effects are zero. In that case, even without any selective publication or manipulation of findings, only 5 percent of significant findings would "replicate." Suppose, alternatively, that all true effects are very large. In that case, almost all replications of significant findings would turn out significant again, no matter how selective the publication process is.

Estimating the Form and Magnitude of Selection

In Andrews and Kasy (2019), we develop two alternative methods for identifying and estimating the form and the magnitude of selection in the publication process. Identifying the form and magnitude of selection allows us to assess the magnitude of implied biases and to correct for them in the interpretation of published findings.

I will use the data of Camerer et al. (2016) to provide some intuition for our methods. Camerer et al. (2016) replicated 18 laboratory experiments published in top economics journals in the years 2011 to 2014. Figure 2 plots data from this systematic replication study in different ways. The left figure shows that the

Figure 2
Evidence for Selective Publication in Economics Lab Experiments



Note: Based on data of Camerer et al. (2016), as explained in the text.

distribution of *z*-statistics in the original studies exhibits a jump at the cutoff of 1.96, suggesting the presence of selection based on significance at the 5 percent level.

The second panel in Figure 2 shows (normalized) original and replication estimates. In the absence of selective publication, there should be no systematic difference between originally published estimates and replication estimates, so that flipping the axes in the figure should not systematically change the picture (leaving differences in sample size aside). In particular, we should find that the points plotted are equally likely to lie above the 45-degree line or below. Selective publication, however, breaks this symmetry. Suppose, for instance, that significant findings are ten times more likely to be published than insignificant findings. Then it will be ten times more likely to observe studies with the combination [original is

significant, replication is insignificant] than with the combination [original is insignificant, replication is significant]. This type of pattern is exactly what we find to be the case for the data of Camerer et al. (2016); lab experiments are much more likely to be published if they find significant effects.

In Andrews and Kasy (2019), we propose a model that allows for an arbitrary distribution of true effects across studies and for an arbitrary function mapping *z*-statistics into publication probabilities (as in Figure 1). This model can be nonparametrically identified and estimated using replication data such as those of Camerer et al. (2016). We can therefore learn from the data how much selection there is and what form it takes. To implement this idea in practice, we propose to assume parametric models: for instance, a step function with jumps at conventional significance levels for publication probabilities, and a *t*-distribution, recentered and scaled with unknown degrees of freedom, for the distribution of true effects across studies. The parameters of such a model can be estimated using maximum likelihood.

The second method proposed in Andrews and Kasy (2019) only relies on the original estimates and their standard errors and does not need replication studies. This method is illustrated in the last panel of Figure 2. This method relies on slightly stronger assumptions and builds on the idea of meta-regressions. In the absence of selective publication, estimates for studies with higher standard errors (and thus smaller sample sizes) should be more dispersed. More specifically, if we take estimates from studies with smaller standard errors and add normal noise of the appropriate magnitude, we should recover the distribution of estimates for studies with larger standard errors. Deviations from this prediction again allow us to pin down fully (estimate) the mapping from estimates to publication probabilities. We propose a model that again allows for an arbitrary distribution of true effects across studies and for an arbitrary function mapping *z*-statistics into publication probabilities, but now assume additionally that standard errors are independent of true effects across studies. This model, or a parametric specification thereof, can be estimated using the data of any meta-study which records estimates and standard errors for different studies. Using this approach, we can again learn how much selection there is, and what form it takes. That is, we can learn what the function mapping z-statistics into publication probabilities looks like.[1]

Estimates of selective publication based on systematic replication studies are valid under very weak assumptions. The estimates based on meta-studies, while relying on stronger assumptions, are much more widely applicable. In settings where we could apply both approaches, we found that both methods yield almost identical estimates.

[1] An app implementing this method, which allows you to estimate selection based on a meta-study, can be found at https://maxkasy.github.io/home/metastudy/. The source code for this app is available at https://github.com/maxkasy/MetaStudiesApp.

Possible Objectives for Reforms of the Publication System

Motivated by concerns about publication bias and replicability, a number of current projects, initiatives, and centers are seeking to improve the transparency and reproducibility of research. These initiatives include the project on Reproducibility and Replicability in Science by the National Academy of Science, the Berkeley Initiative for Transparency in the Social Sciences, the Institute for Quantitative Social Science at Harvard, the Meta-Research Innovation Center at Stanford, and Teaching Integrity in Empirical Research, spanning several institutions. The reforms that have been promoted by these initiatives and others include changes in norms (don't put "stars" based on statistical significance in your tables), changes in journal policies (requiring pre-analysis plans for experimental research, accepting papers based on registered reports), and changes in the institutional infrastructure for academic research (journals for null results and journals for replication studies). We will assess these proposals in the next section. But before doing so, it is useful to take a step back and discuss several alternative objectives that we might wish to pursue in reforming statistics education and the academic publication system: validity, relevance, and plausibility. These alternative objectives can have contradictory implications, which complicates the task of evaluating reforms.

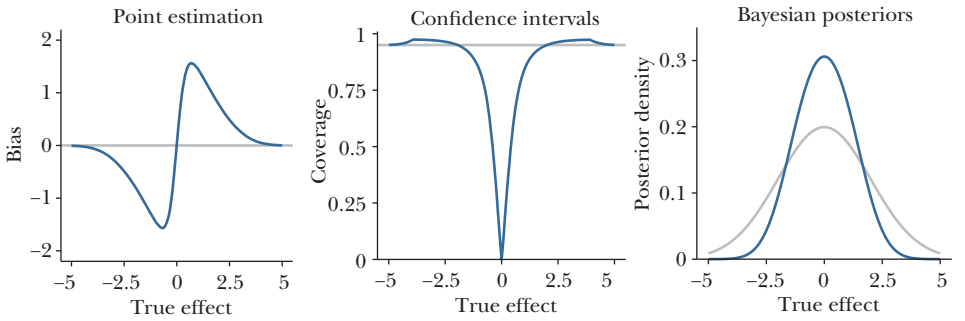
Validity

Why is selection of findings for publication, whether by researchers or by journals, a problem? In canonical settings, standard inference methods are valid if and only if publication probabilities do not depend on findings in any way, although dependence on standard errors is allowed (Frankel and Kasy forthcoming). Any form of selection leads to biased estimates, distortions of size for tests and of coverage for confidence sets, and incorrect Bayesian posteriors—if not properly accounted for.

As an illustration, consider the extreme case where only findings exceeding the 5 percent significance threshold of a z-score of 1.96 (for standard normal estimates) are published. Figure 3 illustrates this case. Each panel in this figure shows the baseline absent selection as a light grey line, and the case of selection as a darker blue line. The first panel shows the bias of point estimates as a function of the true effect, conditional on publication. For very large true effects (whether positive or negative), no bias occurs, because such studies are published with very high probability. For a true effect of zero, no bias occurs either, because positive and negative results are equally likely to be selected. For intermediate effect sizes where the true effect is around 1 standard error, however, point estimates are biased upward by up to 1.5 standard errors from the true value, conditional on publication. This is because studies are only published (in this example) when the estimate exceeds the 5 percent significance threshold.

The middle panel similarly plots the probability that a nominal 95 percent confidence interval contains the true effect, conditional again on the size of the true effect and under the assumption that only results significant at the 5 percent

Figure 3
Distortions Induced by Selective Publication Based on Statistical Significance



Note: The first two plots show the effect of only publishing significant estimates (with a *z*-statistic above 1.96) on the bias of point estimates (average estimate minus truth) and the coverage of confidence intervals (probability of containing the truth) conditional on publication. The third plot shows the effect on the posterior absent publication.

level are published. Again, for large true effects, no distortions happen. When the true effect is small, however, the probability that the confidence interval contains the true effect is much smaller than 95 percent.

Finally, consider a Bayesian reader of the published literature. This reader will update prior beliefs based on the published findings. When observing a published finding, the reader actually does not need to take into account selective publication based on findings. But the reader needs to update beliefs in the absence of a publication! Not observing a publication makes it more likely that the true effect is close to zero, in our example. The last panel of Figure 3 shows two posterior distributions for a Bayesian who starts with a normal prior, when no finding is published (the normal prior is chosen purely for illustration; similar arguments hold for any prior distribution). Relative to the naïve posterior, which ignores selection, the correct posterior that takes selection into account will recognize that the presence of unpublished and unobserved research makes it more likely that the true effect is close to zero, because the Bayesian interprets published findings as a censored sample.[2]

To summarize, there is ample evidence that publication is selective, albeit to different degrees and in different ways across various empirical fields. Selective publication can heavily distort statistical inference, whether frequentist or Bayesian. However, validity of inference should not be the only goal of statistical research. Presumably, researchers also care about ultimate objectives such as scientific progress, social learning, or helping decision-makers in medicine, public policy, and technology. To put it starkly, publishing only estimates calculated based on a random

[2]Alternatively, we could condition on the number of published findings, leading to a truncation-based perspective with very similar implications for Bayesian inference (Andrews and Kasy 2019).

number generator can yield statistical inference that is valid, but completely useless to decision-makers or substantive researchers.

Relevance for Decision-Making

Consider, as an example, that many new therapies in some hypothetical area of medicine—say drugs or surgical methods—are tested in clinical studies. Suppose that most of these trials don't work out and the new therapies just don't deliver. Absent a publication of successful clinical research, no doctor would implement these new therapies. In addition, doctors have limited time: no human can read hundreds of studies every month. But which subset of studies should doctors read? In order to improve medical practice, it would arguably be best to tell doctors about the small subset of new therapies that were successful in clinical trials. In Frankel and Kasy (forthcoming), we derive optimal publication rules when the goal of publication is to inform decision-makers, as in this example. These optimal publication rules confirm the intuition that findings are most useful for decision-makers when they are surprising, and surprising findings should thus have priority in publication.

However, if the selection rule for publication is based on success in a clinical trial, then published findings are biased upward. Replications of the published clinical trials will systematically find smaller positive effects or even sometimes negative effects. This reasoning suggests that there is a deep tension between relevance for decision-making and replicability in the design of publication rules. In Frankel and Kasy (forthcoming), we argue that this type of logic holds more generally in any setting where published research informs decision-makers and there is some cost which prevents us from communicating all the data. Such a cost clearly must be present; otherwise it would be optimal to simply publish all data, without any role for statistical inference, researchers, or journals. Given such a cost, it is not worthwhile to publish "null results"—that is results that do not change decisions relative to the default absent publication. Surprising results, on the other hand, especially those that lead to large changes of optimal decisions, are of great value to decision-makers, and should thus be preferred for publication. This conclusion holds whether or not readers are sophisticated in their interpretation of selectively published findings.

Furthermore, some notions of social learning, such as reducing the variance of posterior beliefs, are isomorphic to the goal of informing decision-makers. Therefore, similar conclusions hold when our goal is to maximize social learning, subject to attention constraints.

Plausibility

Validity of standard inference requires that we eliminate selection on findings, while (policy) relevance encourages us to publish surprising findings. But what about the plausibility of findings? After all, extreme or surprising findings may just indicate that there is some problem with the study design. If a study reports that a very minor intervention has major health benefits, it might be more likely that the reported findings are biased than that the authors stumbled upon a miracle cure.

We can formalize this idea by assuming that readers have some prior distribution over the bias of a study, that is, some prior probability that the study design is flawed. Very surprising findings make it more likely that the bias is large. Very surprising findings therefore lead to less updating of beliefs about the true effect relative to moderate findings.

Suppose now that we are again interested in the relevance of findings for decision-makers. As before, unsurprising findings are not relevant for decision-makers and should not be published. But very surprising findings are implausible, suggesting issues with the study, and should also not be published. Under this model, only intermediate findings satisfy the requirements of both relevance and plausibility.

These considerations leave us with the practical question of what to do about the publication system. How shall we trade off these conflicting objectives? Can we have validity, relevance, and plausibility at the same time? As argued below, a possible solution might be based on a functional differentiation of publication outlets, which could build on the present landscape, while making the differences of objectives and implied publication policies across outlets more explicit. Such a differentiation avoids having to sacrifice one of these objectives (like relevance) for the sake of another (like validity and replicability). But before we get there, let us discuss some specific reform proposals, while keeping in mind the tension between these objectives.

Specific Reform Proposals

Deemphasizing Statistical Significance

Much of traditional statistics—including teaching, editorial guidelines, and statistical software—focuses on the notion of statistical significance. However, a number of academic journals have recently changed their guidance to de-emphasize statistical significance. For example, the American Economic Association advises prospective authors: "Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses." [3]

Debates over the notions of statistical testing and statistical significance have a long history, which we will not recapitulate here. (A companion paper in this symposium by Guido Imbens reviews some issues in these debates.) But for present purposes, it is useful to disentangle four distinct aspects of the common emphasis on testing whether some effect or coefficient is different from a null effect of zero at the 5 percent statistical significance level, before returning to the question of selective publication.

First, there is the emphasis on the largely arbitrary null hypothesis that the true value equals zero, when evaluating estimated results. Arguably, very few effects in the social and life sciences (perhaps in contrast to physics) are exactly equal to

[3] At https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide, accessed January 19, 2021.

zero. For this reason, rejecting the null hypothesis of zero is thus largely a matter of sample size in most applications; with large enough samples, the null hypothesis will always be rejected, because it is wrong. Switching the emphasis of teaching and publishing from significance tests to confidence sets allows us to move away from the focus on this arbitrary value, while maintaining an easily communicable measure of statistical precision.

Second, the 5 percent cutoff for statistical significance is arbitrary, and there is little reason to assume that this cutoff provides a good tradeoff between size and power, that is, between type I errors and type II errors. Reporting point estimates and standard errors, as per the AEA guidelines, provides a resolution to this issue. Point estimates and standard errors are sufficient statistics for the parameter of interest under conventional normal approximations, so that all the relevant information is communicated. In practice, of course, readers trained to think in terms of significance testing might still calculate a test (in their head), comparing estimates to twice their standard error, thus undoing the effect of the reformed reporting standards.

Third, statistical testing imposes a binary interpretation of the data. Empirical research is often discussed in terms of whether the authors "found an effect of X on Y" or not. This is a very coarse representation of data that are usually quite complex. Nothing prevents, in principle, less coarse representations, such as point estimates and standard errors, except perhaps that the latter are harder to summarize or remember. However, the fact that such coarse representations are popular seems to point to attention constraints, which provide one of the motivations for optimal selection rules as discussed in Frankel and Kasy (forthcoming) and in related work by Andrews and Shapiro (2019). Statistical recommendations should take such attention constraints into account.

Fourth, the focus on statistical significance is a major factor driving selective publication, motivated by the notion that effects that are significantly different from zero are somehow more interesting than those that are not. Selection on significance bears some resemblance to selection on surprisingness, which matters for relevance or learning objectives (as discussed above). But neither selection centered at zero nor selection at the 5 percent significance cutoff are optimal for relevance, and they lead to distortions of inference. Selection based on significance should thus be avoided.

Motivated by the observation that very few effects in economics are exactly equal to zero and more generally that few theories can be assumed to hold exactly, Fessler and Kasy (2019) propose an alternative use of economic theory in empirical research that does not involve conventional statistical testing. Instead, we suggest a framework for the construction of estimators which perform particularly well when the empirical implications of a theory under consideration are approximately correct. Our proposed estimators "shrink" empirical findings towards the predictions of a theory. As an example, we might shrink estimated demand functions toward the theoretical prediction that compensated own-price elasticities of demand are negative. By choosing the amount of shrinkage in a data-dependent

manner, we can construct estimators that perform uniformly well and have large gains in performance when the theoretical predictions are approximately correct.

Pre-analysis Plans

Pre-analysis plans have increasingly become a precondition for the publication of experimental research in economics, for both field experiments and lab experiments. Historically, economics first imported randomized controlled trials as a method of choice from clinical research, and then a few years later again followed clinical research (for comparison, see the guidelines from the Food and Drug Administration 1998) in an emphasis on pre-analysis plans. This change in methodological norms has not gone uncontested; for a discussion of the costs and benefits of pre-analysis plans in experimental economics, see Coffman and Niederle (2015) and Olken (2015) in this journal, as well as Banerjee et al. (2020).

In their ideal form, pre-analysis plans specify a full mapping from data to what statistics will be reported. In practice, however, pre-analysis plans often do not specify a full mapping from data to reported results, but instead constrain the analysis and the results to be reported. By tying the researcher's hands, pre-analysis plans prevent the researcher from cherry-picking which results to report. They might thus provide a remedy for the distortions introduced by unacknowledged multiple hypothesis testing. Pre-analysis plans arguably play the same role to frequentist notions of bias and size control as randomized controlled trials play to causality—they are necessary for the very definition of these notions.[4]

In ongoing research (Kasy and Spiess 2021), we take a slightly different perspective. Rather than motivating pre-analysis plans in terms of frequentist hypothesis testing, we propose to model statistical inference as a mechanism design problem. To motivate this approach, note that in pure statistical decision theory there is no need for pre-analysis plans. Rational decision-makers have consistent preferences over time, and thus, no need for the commitment device that is provided by a pre-analysis plan. The situation is different, however, when there are multiple agents with conflicting interests. As an example, consider the conflict of interest between pharmaceutical companies that want to sell drugs and the Food and Drug Administration that wants to protect patient health. Another example would be researchers who want to get published (in order to get tenure) and readers of research who want to learn the truth about economic phenomena.

The mechanism design approach proposed in Kasy and Spiess (2021) takes the perspective of a reader of empirical research who wants to implement a statistical decision rule. Not all rules are implementable, however, when researchers have divergent interests and private information. We characterize implementable rules under these constraints and consider the problem of finding optimal statistical decision rules subject to implementability. In such models, there is a role for

[4]Andrew Gelman makes this point succinctly in https://statmodeling.stat.columbia.edu/2017/03/09/preregistration-like-random-sampling-controlled-experimentation/.

pre-analysis plans under some conditions. In particular, if researchers have many choices (degrees of freedom) for their analysis—there are many forking paths—and if communication costs are high (there is a lot of private information), then pre-analysis plans can improve the welfare (statistical risk) of readers. If, on the other hand, researchers face a smaller number of choices and private information is limited, the reader might be better off without requiring a pre-analysis plan.

Pre-results Journal Review

Pre-analysis plans, at least in theory, eliminate selective reporting of findings by researchers themselves. But they do not eliminate selective publication of findings by journals. In an attempt at eliminating the latter, some outlets such as the *Journal of Development Economics* now allow for submission of "registered reports," where studies are approved for publication based on a pre-results review.[5]

Pre-results review is the policy that most fully implements publication decisions that do not depend on findings but possibly depend on the sample size, question, method, and so on. Such independence of publication from findings is required if our goal is validity of conventional inference. However, such independence is not necessarily desirable if our objective also includes other criteria, such as relevance and plausibility.

Journals for Null Results and Replication Studies

Another recent set of innovations in the publication system are journals dedicated explicitly to null results or to replication studies. Such journals are made possible, in particular, by the reductions in publication cost that come with online-only publication. Economics, for instance, has the *Series of Unsurprising Results in Economics*. Such an outlet, focused on unsurprising or insignificant findings, has a useful role to play in a functionally differentiated publication system. It provides a completion of the record of published findings that can serve as an input for meta-studies and related exercises. There is also the *International Journal for ReViews in Empirical Economics* (IREE), a journal focused on replication studies.[6] Again, replications—with the key caveat of being published independent of findings—can provide a useful addition to a differentiated publication system.

Among other roles, such replications allow for a credible assessment of the selectivity of published findings in some subfield, using for instance the methods of Andrews and Kasy (2019). Extrapolation of estimated selectivity to other findings in the same field then allows for bias corrections in the interpretation of these findings. In addition to allowing us to assess selectivity, replications might also shed light on effect heterogeneity not captured by standard errors, thus providing insight into the external validity of published estimates.

[5]For instance, see https://www.elsevier.com/__data/promis_misc/JDE_RR_Author_Guidelines.pdf, accessed January 19, 2021.
[6]For instance, see https://www.iree.eu/aims-and-scope, accessed January 19, 2021.

Achieving Multiple Objectives in a Functionally Differentiated Publication System

Above, we have argued that alternative objectives—relevance for decision-makers, statistical validity, plausibility of published findings—can lead to conflicting recommendations for reforms of the publication system. However, we might reconcile these objectives by striving for a functional differentiation of publication outlets. The following provides a sketch of such a landscape.

There might be a set of top outlets focused on publishing surprising ("relevant") findings, subject to careful quality vetting by referees. These outlets would have the role of communicating relevant findings to attention-constrained readers (researchers and decision-makers). A key feature of these outlets would be that their results are biased by virtue of being selected based on surprisingness. In fact, this is likely to be true for prominent outlets today, as well. Readers should be aware that this is the case: "Don't take findings published in top outlets at face value."

There might then be another wider set of outlets that are not supposed to select on findings but have similar quality vetting as the top outlets, thus focusing on validity and replicability. For experimental studies, pre-analysis plans and registered reports (results-blind review) might serve as institutional safeguards to ensure the absence of selectivity by both researchers and journals. Journals that explicitly invite submission of "null results" might be an important part of this tier of outlets. This wider set of outlets would serve as a repository of available vetted research and would not be subject to the biases induced by the selectivity of top outlets. Hiring and promotion decisions should take care to give similar weight to this wider set of publications as to top publications, so as to minimize the incentives for researchers to distort findings, whether by p-hacking or other means.

To make the findings from this wider set of publications available to attention-constrained decision-makers, systematic efforts at aggregating findings in review articles and meta-studies by independent researchers would be of great value (Vivalt 2019; Meager 2019). Lastly, systematic replication studies can serve as a corrective for the biases of top publications and as a further safeguard to check for the presence of selectivity among non-top publications.

Summary and Conclusion

Published research is selected through a process that includes both researchers and journals, so that consumers of such research cannot, in general, assume that reported estimates are unbiased, either in their point estimates or their confidence intervals. In this essay, I have argued that conventional methods to detect publication bias have their limitations, but we can identify and estimate the form and magnitude of selection, using either replication studies or meta-studies. I have further argued that replicability and validity of inference should not be our only goal and reform efforts focused on this goal alone are misguided. However, there is a fundamental tension between alternative objectives such as validity, relevance, plausibility, and

replicability. One approach to resolving this tension, at least partially, is to build a functionally differentiated publication system.

Let us conclude by taking a step back to consider what the debates around replicability and selective publication imply for the foundations of statistics. One of the main foundations of statistics is statistical decision theory. The activity of statistics as conceived by decision theory is a rather solitary affair. There is just the researcher and the data, and the researcher has to make some decision based on the data: estimate a parameter, test a hypothesis, and so on. This perspective can be extremely useful. It forces us to be explicit about our objective, the action space, and what prior information we wish to incorporate (for example, in terms of the statistical model chosen, or in terms of a Bayesian prior, or in terms of a set of parameters for which we wish to control worst-case risk). The decision-theory perspective makes explicit the tradeoffs involved in the choice of any statistical procedure.

But this decision-theory perspective also has severe limitations, as evidenced by the discussions around *p*-hacking, publication bias, and pre-analysis plans. It is hard to make sense of these discussions from the vantage point of decision theory. For instance, why don't we simply communicate all the data to the readers of research? If we took decision theory literally, that would be optimal. After all, communicating all the data avoids any issues of selection as well as any waste of information. In practice, as consumers of research, we of course do prefer to read summaries of findings ("X has a big effect on Y, when W holds"), rather than staring at large unprocessed datasets. There is a role for researchers who carefully construct such summaries for readers. But it is hard to make sense of such a role for researchers unless we think of statistics as communication and unless there is some constraint on the attention or time or information-processing capacity of readers.

Relatedly, what is the point of pre-analysis plans? Their purpose is often discussed in terms of the "garden of forking paths" of specification searching. But taking the perspective of decision theory literally again, there is no obvious role for publicly committing to a pre-analysis plan in order to resolve this issue. Researchers might just communicate how they mapped data to statistics at the time of publication. To rationalize publicly registered pre-analysis plans, we again need to consider the social dimension of research; in ongoing work (Kasy and Spiess 2021) we do so through the lens of mechanism design.

These examples illustrate that statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. Taking this social dimension seriously suggests a perspective on statistics where the task of empirical researchers is to provide useful summaries of complex data to their readers in order to promote some form of collective learning. This task is subject to costs of time and attention of researchers, referees, and readers as well as constraints on social learning in terms of limited information, strategic behavior, the social norms of research, and other factors. Elaborating this perspective in which statistics gives normative recommendations for empirical practice, while taking social constraints into account, is an exciting

task for the years ahead. This endeavor will have to draw on a combination of microeconomic theory, psychology, and the sociology and history of science.

■ *This research was funded in part by the Alfred P. Sloan Foundation (under the grant "Publication bias and specification searching. Identification, correction, and reform proposals").*

References

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Andrews, Isaiah, and Jesse M. Shapiro. 2019. "A Model of Scientific Communication." https://scholar.harvard.edu/files/iandrews/files/audience.pdf.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics." NBER Working Paper 26993.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Card, David, and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review* 85 (2): 238–43.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315 (7109): 629–34.

Fessler, Pirmin, and Maximilian Kasy. 2019. "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions." *The Review of Economics and Statistics* 101 (4): 681–98.

Food and Drug Administration. 1998. *Guidance for Industry: Statistical Principles for Clinical Trials.* Rockville, MD: US Department of Health and Human Services.

Frankel, Alexander, and Maximilian Kasy. Forthcoming. "Which Findings Should Be Published?" *American Economic Journal: Microeconomics.*

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis Was Posited ahead of Time." http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8).

Kasy, Maximilian, and Jann Spiess. 2021. "Pre-analysis Plans and Mechanism Design." Unpublished.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11 (1): 57–91.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

Vivalt, Eva. 2019. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 193–214

Evidence on Research Transparency in Economics

Edward Miguel

Open science and research transparency (terms I'll use interchangeably) are advanced when scientific claims are independently verifiable, including through the promotion of free and open sharing of the process of conducting research, and when the content and findings generated during research are objects that readers can check for themselves. A decade ago, "research transparency" and "open science" were not on the radar screen of most economists or other social scientists. However, a new scholarly movement has coalesced around bringing new open-science practices, tools, and norms into the mainstream. Prominent social science organizations have taken the field, including the Center for Open Science (cos.io), the Society for the Improvement of Psychological Science (improvingpsych.org), and the Berkeley Initiative for Transparency in the Social Sciences (bitss.org). The goal of this article is to lay out the emerging evidence on the adoption of these approaches in three specific areas—open data, pre-registration and pre-analysis plans, and journal policies—and more tentatively, to begin to assess their impacts on the quality and credibility of economics research.

In a broad normative perspective, the open science movement seeks to bring the research practices and culture of economics in line with a classical "scientific ethos" of open inquiry that goes back centuries. In one prominent discussion, Merton (1942) laid out the four so-called "Mertonian norms" of scientific inquiry: universalism, communality, disinterestedness, and organized skepticism. In the

■ *Edward Miguel is the Oxfam Professor of Environmental and Resource Economics, University of California Berkeley, Berkeley, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is emiguel@berkeley.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.193.

social sciences, one aspect of these norms is the basic ability of others in the scholarly community to reproduce published findings, and thus, to understand fully how they were produced and what alternative analyses might be possible. In economics, relatively few papers—in the range of one-third to one-half of published articles (Chang and Li 2015; Gertler, Galiani, and Romero 2018)—achieve even the basic goal of computational reproducibility, which involves using the publicly shared materials to produce the results in the paper. Much of this has to do with the quality of the replication data and code, which are often incomplete or poorly documented.

In addition, the movement for open science and research transparency seeks to reduce the extent to which investigator bias or other biases can creep into both research practices and publication decisions. There is ample evidence of pervasive problems in empirical economics research (and in related quantitative fields like political science and social psychology). I won't present a comprehensive account of these concerns here—for more detail, readers can refer to Christensen and Miguel (2018) and the Christensen, Freese, and Miguel (2019) book—but it is worth offering some concrete examples.

One common pattern is that if a study produces a "null result" that is not different from zero at a conventional level of statistical significance, typically 5 percent, it is less likely to be written up by a social science researcher and less likely to be published if it is written up (for example, Franco, Malhotra, and Simonovitz 2014). High proportions of null results "disappearing" have been documented in well-designed studies in economics (Andrews and Kasy 2019) and other fields (Turner et al. 2008; Simonsohn, Nelson, and Simons 2014). When null results disappear from public view, these unwritten findings are effectively lost to the broader research community. The result is that looking at published research may lead to misleading conclusions regarding topics of intellectual and public policy importance (Ioannidis 2005). The disappearance of nulls also wastes research funding and human resources by producing efforts that are never published and leads to duplicated efforts when other scholars carry out work that (unbeknownst to them) was already tried earlier.

This issue of the disappearing nulls is closely related to *publication bias*, and the related concern about selective presentation of results. If authors believe that results need to attain an arbitrary level of statistical significance to be meaningful or publishable, they will have an incentive to manipulate their research accordingly, in what is sometimes called *p*-hacking or phishing. Empirical economists have traditionally engaged in lengthy periods of largely unstructured data-mining, in which potentially thousands of statistical tests were run, but they then only report their handful of "preferred" estimates in the final manuscript (Leamer 1983). Such choices lead almost unavoidably to cherry-picking of results to obtain p-values below 0.05 and inflated statistical significance levels. Brodeur et al. (2016) and Brodeur, Cook, and Heyes (2020) demonstrate in leading economics outlets, including top-five journals, that there are substantial spikes in empirical estimates barely above the statistical significance level of 5 percent, with apparent "canyons" just below this value—a clear sign that the published research was pre-selected by authors or journals to meet this standard. Gerber and Malhotra (2008a, 2008b) show a similar pattern of empirical results in leading political science and sociology journals.

In the discussion below, I present evidence indicating that economics is in a period of rapid transition toward new transparency norms in the areas of open data, preregistration and pre-analysis plans, and journal policies.[1] I will argue that there are indications of at least some social benefits from these practices. There is also growing reason to believe that critics' worst fears regarding their potential costs—like onerous adoption costs or the stifling of creativity—have not been realized. I close by arguing that further cultural change is needed to reinforce and sustain the changes that are already underway in economics.

Open Data

When I was a graduate student in the 1990s, obtaining the underlying data and analysis scripts for a published paper was typically either challenging or impossible. But dramatic changes in technology—especially the rise of the internet over the past 25 years—and in policies of academic journals and professional associations have led rapid shifts in prevailing practices.

In economics, one catalyst for these changes was the data-sharing policy adopted by the American Economic Association (AEA) in 2005, which came in response to growing evidence that many, if not most, published empirical analysis in economics could not be readily reproduced (Dewald, Thursby, and Anderson 1986; Bernanke 2004). The policy led to an almost immediate increase in the posting of data and analysis code for the *American Economic Review* (Christensen, Dafoe et al. 2019). Other leading general interest journals and field journals followed suit, with many adopting the AEA policy verbatim (Christensen and Miguel 2018). This has led to a dramatic increase in access to data for published research in our discipline in a relatively short span of time.

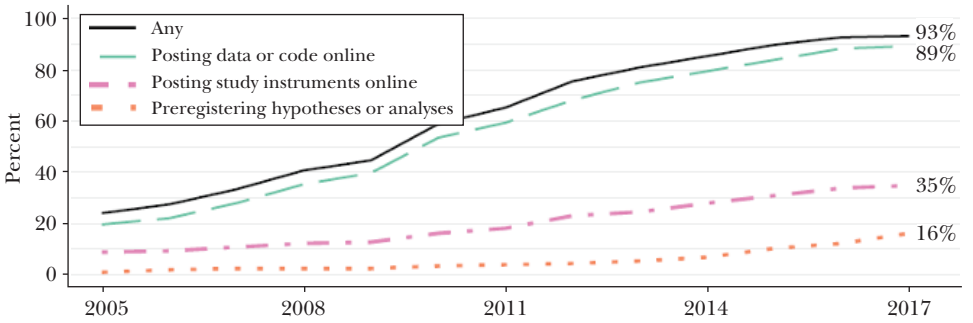
Figure 1 illustrates the rise of data sharing in economics, based on an attempt to obtain a representative survey sample of scholars who had published in top-ten economics journals during 2014–2016 (as well as in other social science fields, although here we discuss economics).[2] We achieved a respectable response rate of nearly 50 percent.[3] While relatively few of the economists in our sample had shared data circa 2005 (just before the AEA policy was adopted), by 2017 nearly 90 percent

[1]Another open science practice that has gained traction in economics is *disclosure*, including the 2012 AEA policy mandating that authors disclose personal, professional, and financial conflicts of interest (https://www.aeaweb.org/journals/policies/disclosure-policy). While I briefly mention the importance of disclosure below, I do not focus on it here: for additional discussion, see Miguel et al. (2014). A further research transparency practice that has long been important in other fields, especially medical research (through the CONSORT guidelines, Begg et al. 1996), is the establishment of author *reporting guidelines*. While promising, this practice has not yet caught on in economics, and again, I do not discuss it here.

[2]Christensen, Wang et al. (2019) present data across disciplines. Patterns over time are similar in economics and political science. Psychology is ahead of economics in the adoption of preregistration but behind economics in data sharing, while sociology has the slowest adoption on all measures. Differences in research methods by field (for example, the prevalence of experimental versus non-experimental studies, and quantitative versus qualitative approaches) appear to explain many of the gaps.

[3]We also restrict the analysis in the figure to scholars who had received their PhD before 2009 (N = 204), meaning they had been active researchers long enough for the time series of their research practices to be meaningful. We also surveyed PhD students, but do not report their behaviors in the figure.

Figure 1
Research Transparency Practices Are Rising in Economics



Source: From Swanson ⊙ al. (2020, Figure 1, panel A).
Note: The chart shows for a given year the proportion of published authors who report having first completed an open science practice in that year or previously. The solid black line shows the proportion of published authors who had completed any open science practice by that year. The dashed green line shows the proportion who had posted data or code online by that year. The dash-dotted purple line shows the proportion of published authors who had posted study instruments online by that year. The dotted orange line shows the proportion who had preregistered an analysis or hypothesis by that year. Posting study instruments online is the response to the question, *"Approximately when was the first time you publicly posted study instruments online?"* Posting data or code online is the response to the question, *"Approximately when was the first time you publicly posted data or code online?"* Preregistering hypotheses or analyses is the response to the question, *"Approximately when was the first time you preregistered hypotheses or analyses in advance of a study?"* The sample is restricted to published authors who completed their PhDs by 2009 (N = 204).

had publicly shared data at least once. There was a similar trend, although with lower levels, for sharing study instruments such as field surveys or lab protocols; there is less adoption of preregistration and pre-analysis plans (which we discuss in the next section) although it too is rising over time.

Many economics journals now directly host data and code on their own website, while there is also growing popularity of internet data repositories, including well-known outlets such as the Harvard Dataverse and Inter-university Consortium for Political and Social Research (ICPSR). These have been so rapidly successful that it is easy to forget what an important innovation the professional curation, storage, and management of research data and code has been.

Yet journal data-sharing policies are not a panacea. The threat of replication alone may provide only weak incentives for academic integrity from scholars who do not anticipate their study will garner extensive interest or citations. Replication materials for published papers are too often poorly documented and disorganized. Partly in response to such concerns, in 2019 the American Economic Association adopted a new and more ambitious set of data sharing standards: the updated AEA Data and Code Availability Policy can be found at https://www.aeaweb.org/journals/policies/data-code. Some key additions are the requirement that data and code be submitted to AEA journals before final paper acceptance and that it be posted on a data repository (openICPSR, rather than at a journal website). There

is also an expanded role for the AEA Data Editor and associated staff, who carry out pre-publication verification of analytical results whenever feasible (Vilhuber 2020), where the Data Editor's role is to assess computational reproducibility, not to judge the appropriateness of the underlying econometric choices. Across AEA journals, the Data Editor's team carried out 216 pre-publication assessments across 138 papers between July and November 2019, and none had "fundamental flaws" (Vilhuber, Turrito, and Welch 2020), meaning any issues found were communicated to the authors and resolved before publication.

It is worth noting several limitations of the AEA data requirements and of open data policies in general. First, not all data can be accessed by the AEA team—for instance, if it is proprietary or subject to government confidentiality restrictions. In these cases, one option is for verification of results by a third-party replicator who has access to the data used in the author's paper; for details, see https://aeadataeditor.github.io/aea-de-guidance/protocol-3rd-party-replication.html. Second, in most cases the data that is shared through the AEA process is not the detailed micro-data, but rather, an aggregated and processed file. Posting the raw underlying data (wherever possible) would generate additional social value for the broader research community, and moves to encourage this would be a useful direction for future open data reforms by the AEA and other economics associations and journals.

However, by conditioning final paper acceptance on including relatively high-quality replication materials across the range of AEA journals, the Association has raised the bar for the field and brings economics close to what is considered "best practice" across other scientific disciplines as captured in the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al. 2015). It seems likely to me that other economics journals will eventually follow suit, as they did after the adoption of the 2005 AEA guidelines. According to the TOP guidelines, which capture open data as well as other practices discussed below, the AEA journals currently rank as the most compliant with open science standards (among the 50 most cited economics journals), with TOP scores similar to leading general scientific journals like *Nature* and *Science* (Bogdanoski and Stillman 2021).

An impressive 97 percent of economists express support for data sharing in the Swanson ⓔ al. (2020) survey data—although respondent beliefs about their colleagues' support for research transparency practices is consistently below respondents' stated support. While most economists believe the rise in data sharing has been worth it, it is still valuable to assess costs, private benefits, and social benefits.

On the cost side, the time it takes to assemble replication materials for a forthcoming article could take anywhere from several hours to a few weeks of work, based on my own personal experience and anecdotally. The variation across projects is related to the size and complexity of the underlying dataset, of course, but is also greatly affected by whether the data, code, and documentation materials are put together along the way as a project is being conducted, or if one needs to assemble them after analysis has been completed. Creating documented data and code materials from a project completed years earlier can be particularly time-consuming and difficult. However, the shift to a new norm of (nearly) universal sharing of data and code means that economists today know they will need to share these materials with

other scholars going forward (for publication in a prestigious AEA journal, say), giving them every incentive to comment generously on their code, label variables clearly, write README files, and generally keep materials organized along the way.

There is some quantitative evidence on the time costs of preparing data and code materials. Since 2016, Innovations for Poverty Action (IPA), a development economics research organization, has funded staff who are tasked with preparing replication materials for field data collection projects that they had supported, and they recorded the time it took (IPA 2020).[4] Across 65 project datasets, the average amount of time to prepare replication materials for public sharing was 31.5 hours, with an interquartile range of 10.0 to 40.5 hours (and a 10th to 90th percentile range of 5.8 to 80.2 hours). This is non-trivial for most projects: still, remember that this estimate of preparation time applies to field experiments that often require multiple years of work on collecting data, so it remains a very small share of overall project work time.

A frequently discussed concern is that enhanced data- and code-sharing requirements will be particularly onerous for scholars who lack the resources to hire a dedicated research assistant, thus exacerbating existing inequalities among researchers. For scholars at an early career stage or who are not working in resource-rich institutions, including many in low- and middle-income countries, these 31.5 hours of work (on average) to assemble data and code for posting will need to be carried out on nights and weekends, given already heavy faculty teaching loads and administrative responsibilities. A promising solution could be for more research funders to dedicate resources to making data and code publicly available, such as efforts recently carried out by IPA, the Berkeley Initiative for Transparency in the Social Sciences (BITSS), and the Jameel Poverty Action Lab (J-PAL). Beyond providing a fairer playing field for all economists, expanded funding for dataset and code preparation would help align private and social incentives for the creation of these research public goods.

The most immediate private benefit that I and many other scholars have personally experienced from new open data norms is the fact that our own research data is better organized *for ourselves* and thus easier to reuse for other analyses and papers as a result of the effort that is put into improved documentation (like the README files and other replication materials). Many scholars (myself included) have often procrastinated on assembling data documentation materials and doing the final grunt work needed to get a dataset ready for sharing with other scholars. Journal policies that force one to do this to get your paper to final acceptance and publication do focus the mind.

Another private benefit from public data sharing is the possibility that it will lead to further related work by others, and thus to greater citations and influence. The likelihood that data-sharing will generate citations has been enhanced by the policies of the AEA, other journals, and nearly all data repositories to provide digital object identifiers (DOIs) for posted research datasets. A number of scholars have shown that data sharing at the article level is correlated with higher citations for that

[4] I am grateful to Hasina Badani of the IPA Research Transparency, Data Governance and Ethics Team for generously sharing this data.

paper (for instance, Piwowar, Day, and Fridsma 2007; Piwowar and Vision 2013), although there remain obvious omitted variable concerns associated with the nonrandomness of the data-sharing decision.

In Christensen, Dafoe et al. (2019), we attempt to address the possible selection into data-sharing using the 2005 AEA data-sharing policy as a natural experiment. In particular, we compare papers published in the *AER* versus the *Quarterly Journal of Economics (QJE)* for four years before and after the 2005 policy change. The availability of data and code for *AER* articles increased quickly, while rates of data availability at the *QJE* (which did not adopt a comparable data policy until 2016) remained low in our study period. In addition, average article citations (through November 2017) rose roughly 50 percent for articles published in the *AER* after the policy change. These results should be viewed as provocative, rather than definitive, given the sample of two journals. Yet the possibility that posting data and code generates higher citations would create strong private incentives to support open data and may help to explain why open data has quickly become a strongly held norm in economics.

One possible social benefit is that open data may enable other scholars to uncover research fraud more readily: for example, discoveries of fraud in political science and social psychology were enabled by journal open data policies. In one vivid example, Broockman, Kalla, and Aranow (2015) downloaded replication data and code from the website of *Science* and discovered statistical anomalies, including too little variation in key measures, which they correctly concluded were consistent with the data having been generated by a random number generator rather than collected in the field. The strength of open data norms in economics—which emerged several years before other social sciences (Christensen, Wang et al. 2019)—could partially explain why there have been fewer high-profile instances of research fraud in our discipline in recent years.

Perhaps the most widely discussed potential social benefit of open data is the opportunity it provides for other scholars to gain a deeper understanding of the research and to build from it. For example, a reanalysis can consider the robustness of findings: that is, do the findings change substantially with modest changes to the specification or research approach? Replication can apply the same research method to a different dataset. More broadly, earlier results can be extended by looking at the results of variations in the underlying model, data over a longer time period, and so on. In this way, embracing research transparency can also be a step toward a fairer and more inclusive scholarly community. I believe that making replication of empirical analysis the norm will have major scientific benefits for economics in the long-run. Indeed, as Maximilian Kasy explains in his paper in this symposium, findings of later replications can allow other scholars to quantify the extent of publication bias in economics, together with associated econometric approaches to correct for it.

However, researchers' growing ability to access data and code from previous studies has led to some controversy (for discussion, see Christensen and Miguel 2018). Gertler, Galiani, and Romero (2018) note that there may be "overturn bias," in which reanalysis and replications that contradict an originally published paper are seen as more publishable. It will be important to address this incentive, in part

by making sure that "successful" replications are also published, not just those that claim to debunk earlier findings. Other replicating authors may also be motivated by personal or monetary conflicts of interest—for instance, if their work is funded by a research sponsor with a financial stake in the answer to the question at hand, such as a pharmaceutical firm or energy company—making strong conflict-of-interest disclosure requirements even more essential.

But ultimately, in my view, many of the concerns around replications from greater openness of data and code are growing pains due to the fact that we are in a transitional phase between an earlier era (like my grad school days) when data was rarely available and replications seldom carried out, and a not-too-distant future when reproducibility of results and other data checks like those in the AEA journals will have become *de rigueur* for both researchers and journals.

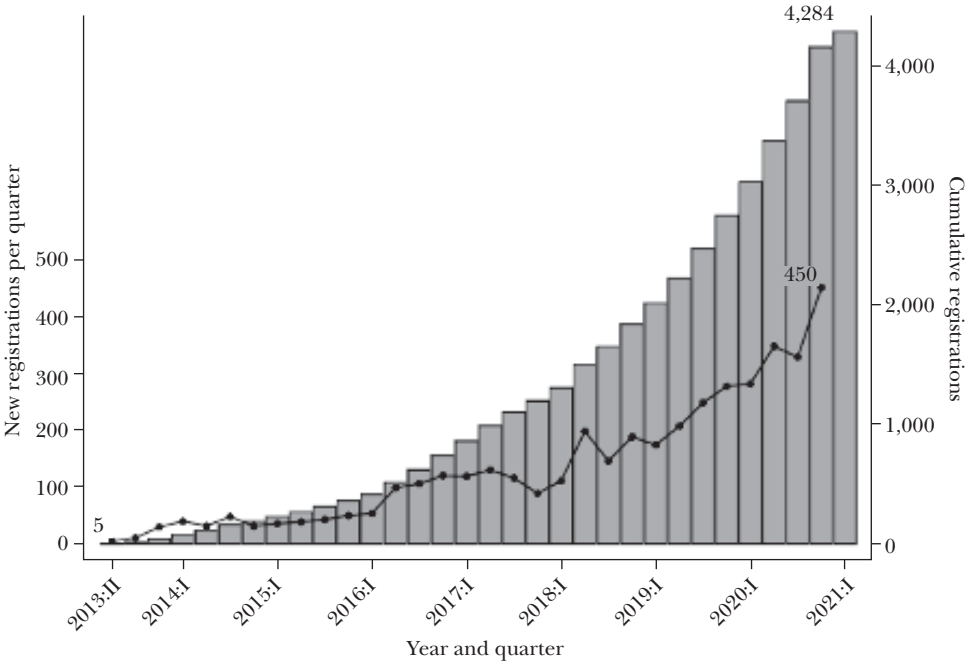
Stepping back, perhaps the most important lingering concern about the expansion of data-sharing requirements is the potential for reduced incentives to collect new data. As Christensen Freese, and Miguel (2019) note: "Data are the lifeblood of empirical science, and it would be a perverse consequence of a data-sharing policy if it reduced the amount of important data collected." There is clearly a need to balance these incentives for the generation of new data versus the social gains of greater sharing of such data, and to do so approaches like temporary "data embargoes" (similar in spirit to technology patents) could be useful. Continuing the current norm of time-limited monopoly rights over the use of data that scholars have generated themselves could be essential to incentivize researchers to carry out ambitious future data collection projects. More thought and debate are still needed regarding how to strike the right balance between these competing concerns to craft the most effective data-sharing policies in economics; in doing so, it will be useful to learn from the experiences of other scientific fields (Hill, Stein, and Williams 2020).

Preregistration and Pre-analysis Plans

Among the open science innovations that have taken place in economics over the last 15 years, the creation of a study registry and growing use of pre-analysis plans is arguably the biggest break with previous research practices. Since its founding in 2013, the AEA Randomized Control Trial Registry has seen exponential growth; by January 2021, over 4,200 studies were registered, as shown in Figure 2. The registry asks for basic study characteristics, like the where, when, and what of the data, approval of an Institutional Review Board, and a few other items. Since 2017, 45 percent of newly registered prospective studies have also posted a pre-analysis plan, with additional more detailed step-by-step description of just how the analysis will be carried out.[5] Similar changes are underway in other social sciences: in political science, the Experiments in Government and Politics (EGAP) registry is widely

[5]This statistic is based on publicly available data downloaded from the AEA registry (https://www.socialscienceregistry.org/) on January 31, 2021; see https://doi.org/10.7910/DVN/FUO7FC.

Figure 2
Studies Posted over Time, American Economic Association Randomized Controlled Trial Registry



Source: This figure was produced by Garret Christensen, Edward Miguel, and Sarah Stillman, and is in the public domain at https://doi.org/10.7910/DVN/FUO7FC. Cumulative and new registrations of studies (by quarter) on the AEA Registry for Randomized Controlled Trials. Data downloaded on January 31, 2021 from https://www.socialscienceregistry.org/. The quarterly figure is not shown for Quarter 1 of 2021 (since data is only available to date for the first month of that quarter).

used; in psychology, most scholars register either on the Open Science Framework (OSF) or on AsPredicted.

Views towards preregistration and pre-analysis plans are generally positive in economics, but with some doubts. The Swanson ⓔ al. (2020) survey data indicate that slightly over half of economists support these practices (with many expressing indifference); in development economics, the subfield where study registration and pre-analysis plans first took off, stated support is far higher at 80 percent. My goal in this piece is not to rehash the ongoing debates about potential benefits of adopting preregistration, and whether they justify the up-front costs. For an overview of these debates, the reader can turn to Olken (2015) and Coffman and Niederle (2015): for additional views, recent starting points are Christensen, Freese, and Miguel (2019); Duflo et al. (2020); and Abrams, Libgober, and List (2020). Rather I will briefly sketch the parameters of the existing debate and then devote my attention to newly emerging evidence on the real-world practice and implementation.

The case for preregistration and pre-analysis plans comes in a few flavors. First, a registry creates a "paper trail," which can help scholars working in an area to learn

about each other's ongoing work. Second, preregistration and pre-analysis plans generate accountability: the rest of the research community (including journal referees) can see which questions the scholars initially intended to ask and this can help reduce publication bias by increasing the reporting of all results, including null results. The option on the AEA registry to keep pre-analysis plans temporarily private (before a paper with the results is released, for instance) reduces concerns that other scholars will troll the registry to "scoop" particularly innovative ideas. Third, a pre-analysis plan can reduce pressure on researchers to emphasize a certain subset of results that may be favored by government officials, research funders, or even colleagues. Finally, an underappreciated benefit of preregistration and a pre-analysis, in my view, is that it improves the quality of the research by pushing scholars to think more carefully about their design and data beforehand. I return to this point below.

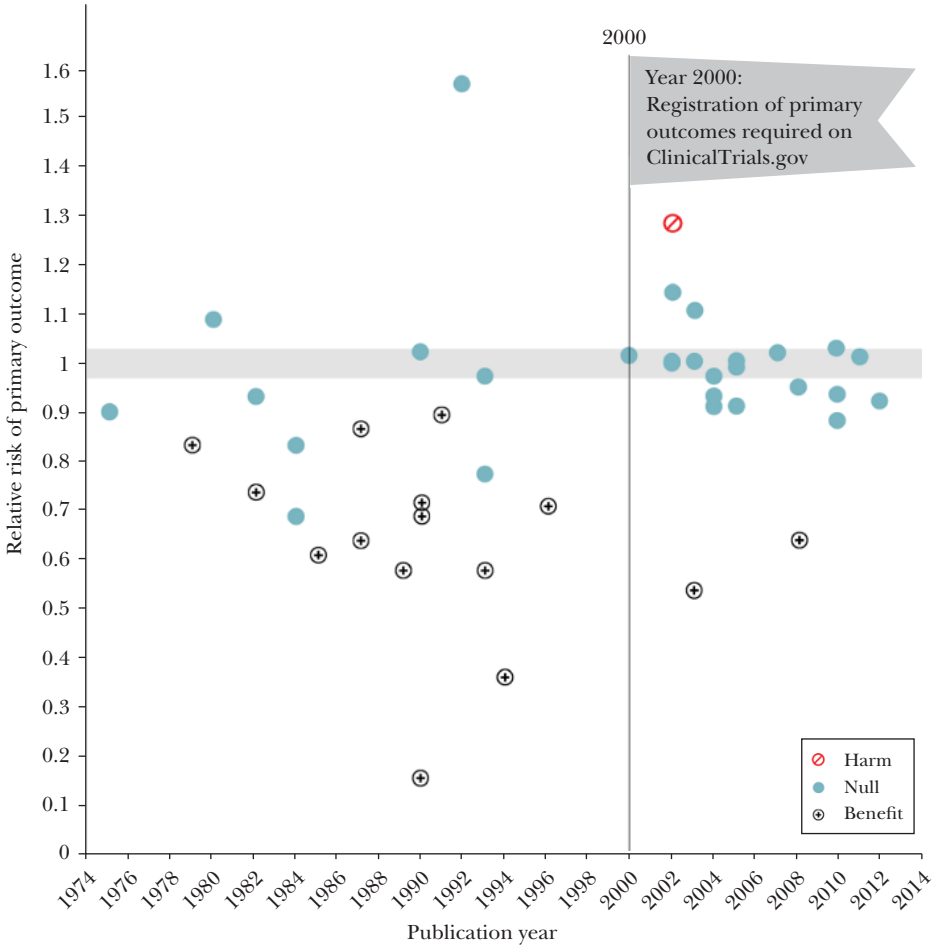
There are also potential costs. First, Olken (2015) mentions the time costs, which in turn will depend on the level of detail needed beyond the basic study characteristics demanded in the AEA registry (as discussed in Duflo et al. 2020). Second, authors fear that pre-imposed constraints on their analytical work may produce an end-product that is less creative and interesting—and possibly less publishable. However, this second concern seems overstated to me. A norm has quickly emerged in economics that allows—and even encourages—authors to present additional analyses to whatever was prespecified, with the caveat being that authors must transparently report what was and wasn't in their plan. Indeed, the first two papers published in economics that employed a pre-analysis plan (to my knowledge), namely, Finkelstein et al. (2012) and Casey, Glennerster, and Miguel (2012), both describe why they felt it was also necessary to publish some analyses that went beyond their pre-analysis plans, and they clearly label these results as such.

While we do not yet know for sure what registration will do in economics in the long-run given how recently the AEA registry was set up, we can learn from the experience of other fields. In particular, the rise of randomized control trials in economics was preceded by the growth of medical trials, and the creation of the AEA registry in economics was directly inspired by ClinicalTrials.gov, which was set up in 2000.

Several benefits have been documented in clinical trial research from having a registry. First, it has become possible to assess how published papers deviate from original plans. A number of studies have audited these deviations in medical studies (such as Mathieu et al. 2009), something that could easily be adopted in economics to immediately provide a greater level of accountability and make sure fewer results disappear.

Second, the creation of a clinical trials registry in medical research appears to lead to more reported null results. In Figure 3, reproduced from Kaplan and Irvin (2015), each dot represents a study on nutritional supplements funded by the same funding arm of the National Institutes of Health; the fact that all were chosen for funding provides some degree of study comparability and quality control. The vertical line marks the founding of ClinicalTrials.gov in 2000. The pre-post research design here is obviously not ideal, but the pattern is striking. Before the registry, the majority of published results were statistically significant and showed benefits, many

Figure 3
Relative Risk of Treatment by Publication Year



Source: Reproduced from Kaplan and Irvin (2015), Figure 1 (Creative Commons Attribution, CC BY, license).
Note: Data are from large NHLBI trials on pharmaceutical and dietary supplement interventions. Positive trials are indicated by the plus signs while trials showing harm are indicated by a diagonal line within a circle. Prior to 2000 when trials were not registered in clinical trials.gov, there was substantial variability in outcome. Following the imposition of the requirement that trials preregister in clinical trials.gov, the relative risk on primary outcomes showed considerably less variability around 1.0.

with large effect estimates. After ClinicalTrials.gov was set up and medical journals began requiring study registration as a publication requirement (De Angelis et al. 2004), far more null results appeared in the literature, and in fact, hardly any significant positive results showed up.

In the decades before 2000, there were repeated scandals in medical research involving clinical trials funded by self-interested pharmaceutical companies, often accompanied by some evidence that "null" trial results that would have hurt these

firms' bottom lines systematically went unreported (Turner et al. 2008). The existence of the trial registry combined with journal requirements to preregister makes this much harder to do, making the clinical trial literature more credible.

Could similar benefits emerge in economics? In economics (and political science), the most detailed evidence to date on the real-world use and impacts of preregistration and pre-analysis plans comes from two papers by Ofosu and Posner (2020a, 2020b). Ofosu and Posner (2020a) examine all working papers released by the National Bureau of Economic Research from 2011 to 2018 and searched for all that used experimental (field and lab) research methods—because these methods are most likely to preregister and write pre-analysis plans. They then search among these working papers for those that also mention a pre-analysis plan. In all, 8.4 percent of experimental working papers during this period mention the existence of an associated pre-analysis plan, with rates rising over time. Ofosu and Posner then determine which of these papers were ultimately published (and where), and through web searches gather total citation counts on Google Scholar as of 2019. They ask whether experimental papers in economics that used a pre-analysis plan have different publication and citation trajectories than those that did not. Of course, adoption of pre-analysis plans was not randomly allocated, but they argue that their focus on the subfield of papers using experimental methods and the fact that all are written by NBER affiliates means they are not comparing apples to oranges. That said, the authors emphasize that the results should be treated as "suggestive" and as a "snapshot."

Ofosu and Posner (2020a) find that the overall likelihood of being published is somewhat lower for studies with a pre-analysis plan (44 percent) compared to those without one (at 54 percent, though this difference is not significant at traditional levels). However, studies with pre-analysis plans are more than twice as likely to have been published in "top-five" economics general interest journals than others (27 percent versus 12 percent).[6] Studies associated with a pre-analysis plan also have 60 percent higher citations by 2019. The authors do not provide a definitive answer for why studies with pre-analysis plans receive more citations and are published in more prestigious journals. One possibility is that perhaps stronger researchers tended to adopt pre-analysis plans sooner or did so for their most promising projects. Another possibility is that studies with pre-analysis plans that obtain a null finding might find it easier to be accepted for journal publication: for example, the first two pre-analysis plan papers in 2012 both contained results that could be seen as "disappointing" or went against some scholars' priors but were still both published in a leading journal (Casey, Glennerster, and Miguel 2012; Finkelstein et al. 2012). Finally, perhaps the process of writing a pre-analysis plan improves the research, leading to stronger papers that are easier to publish in leading venues.

Their second study (Ofosu and Posner 2020b) builds on a novel survey conducted among scholars in economics and political science who belonged to networks specializing in experimental research—and were thus likely to have registered pre-analysis plans—regarding their experiences, practices, and beliefs. They

[6] The authors define the top-five journals in the usual way: *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

also review the content in a representative subset of 195 registered pre-analysis plans. The survey has some limitations. It has a relatively low response rate among those contacted (at 23 percent). Also, it focuses on pre-analysis plans written through 2016, which places this data in the early days of preregistration. Norms may have evolved considerably since then. Still, their data remains among the best available sources of quantitative information (to my knowledge) on real-world use of pre-analysis plans.

For example, the modal time to write a pre-analysis plan is two to four weeks of work time among the survey respondents, a figure that resonates with my own experience. Most survey respondents also mention that this time is not all additive, because it is much faster to move directly into analysis mode if you have already spent weeks carefully laying out the regressions that will be run and thinking through how to avoid certain pitfalls. In fact, 33 percent of respondents indicate "that these [time] savings were equal to or greater than the time spent to draft the PAP in the first place"(Ofosu and Posner 2020b). That said, a pre-analysis plan may impose larger time costs on some scholars, perhaps because some research is intrinsically more complex, or because some researchers tend to write quite detailed pre-analysis plans (myself included) while others focus on a tighter subset of analytical issues (as discussed in Duflo et al. 2020).

The survey evidence also suggests potential quality benefits to writing pre-analysis plans: "An overwhelming majority (8 in 10) said that drafting a PAP caused them to discover things about their project that led to refinements in their research protocols and/or data analysis plans" (Ofosu and Posner 2020b). Ofosu and Posner (2020b) advocate finding ways to harness this potential advantage of the pre-analysis plans by getting early feedback on the research plan before registering it. Indeed, pre-analysis plans are already starting to be incorporated as a normal research product to present in some venues, including the Working Group in African Political Economy (WGAPE) meetings.[7]

Finally, Ofosu and Posner (2020b) assess registered pre-analysis plans along four dimensions: "specifying a clear hypothesis; specifying the primary dependent and independent/treatment variable(s) sufficiently clearly so as to prevent post hoc adjustments; and spelling out the precise statistical model to be tested." Here the record is mixed. In their sample, 90 percent of pre-analysis plans state a clear hypothesis and 80 percent contain at least three of the four elements. However, many of the resulting papers report results that were not in the original pre-analysis plan without always clearly labeling them as such. It remains possible that this situation has improved since their data from 2016, but updated research could document how the use of pre-analysis plans has evolved over time.

Abrams, Libgober, and List (2020) carry out a related audit of the pre-analysis plans listed on the AEA registry. They point out that norms regarding registration vary considerably even across experimental fields, with high rates among economists conducting field experiments but far lower levels among those carrying out lab experiments. They also provide a set of useful reform proposals, including possibly mandating registration before projects are carried out, greater incentives

[7] Dan Posner and I have co-organized WGAPE meetings together with several colleagues since 2002.

to post results generated by the research, and posting materials from Institutional Review Boards.

I cannot claim to have a final answer on whether the benefits of pre-analysis plans exceed their costs, although it seems clear that the more dire predictions from the early days of the AEA registry regarding onerous time costs and stifled creativity have not been borne out. When Ofosu and Posner (2020b) ask directly, 64 percent of scholars respond that "[writing a PAP] takes a considerable amount of time, but it is worth it," while 6 percent write that "It doesn't take much time, so the cost is low," meaning that 70 percent of researchers actively working in this area are largely positive about the benefit to cost ratio. This lines up with the 80 percent of development economists (surveyed in Swanson ⓔ al 2020) who support preregistration.

My sense as a co-author, colleague referee, and adviser is that there is still considerable variation in the style of pre-analysis plans that economists are writing: some are more detailed, others less, some contain more literature review or conceptual discussion, some don't, and so on. My own view is even a relatively sparse pre-analysis plan that lays out the primary outcomes, the core analysis, and main regression specifications remains useful in addressing the most extreme forms of selective reporting and data mining as well as publication bias. Other leading economics associations, including the European Economic Association and the Econometric Society, have moved partially in the same direction regarding registration, and "encourage authors of papers that use RCTs [randomized control trials] to register their experiments" but do not (yet) mandate it.[8]

Preregistration to date has largely been utilized in fields that employ experimental methods, including applied microeconomics fields (especially development economics) and experimental economics. Preregistration and pre-analysis plans have made far less headway in structural econometric work, including in industrial organization, international trade, and macroeconomics. Preregistration appears to be more challenging to implement in structural work, where underlying theoretical models are often more complicated and their construction and estimation involves myriad judgement calls that may be challenging to anticipate and specify in advance—and also more difficult for outside observers to discern. The resulting increase in researcher degrees of freedom likely makes it harder to detect biased reporting. One immediate way forward in these fields—albeit partial—would be for at least some steps of the research process to be prespecified, for instance, the value of particular parameters (like the intertemporal discount rate) to be used in quantitative exercises, or the specific dataset to be analyzed. In the absence of preregistration, a broader discussion is needed in these fields regarding whether there are alternatives that could enhance transparency and similarly constrain *p*-hacking, lest we witness a growing methodological breach across economics subfields over time.

[8]These policies (as of May 31, 2020) are at https://www.eeassoc.org/index.php?site=JEEA&page=42 and https://www.econometricsociety.org/publications/econometrica/information-authors/instructions-submitting-articles, respectively.

Journal Policies and Practices

Journal policies and practices are influential in setting norms in any scientific field. Here, I will assess two policy changes related to open science issues that have recently been implemented at high-profile economics journals: *pre-results review* and *editorial statements*. Behind both policies is the notion that economic research should be judged by authors and journals based on whether the project was worth undertaking in the first place.

Specifically, the idea behind pre-results review is that referees and editors should ideally judge the quality of a research paper based on its design, data, and the importance of the underlying question, rather than being influenced by whether the results are surprising, well-suited for a press release, statistically significant, or confirm (or contradict) prevailing theory. This approach has become more common in other social science fields, especially psychology and cognitive science, where papers published using this approach are often called "registered reports."

One immediate objection to pre-results review might be that scholars lack the capability to evaluate submitted articles without seeing the results. However, scholars evaluate research proposals that lack results all the time: for instance, when sitting on National Science Foundation or National Institute of Health panels that review grant proposals, when deciding which graduate student travel awards to fund, or when serving on a dissertation prospectus committee. Growing familiarity with pre-analysis plans also facilitates pre-results review.

With pre-results review, an empirical article goes through two stages of review During the first stage, authors submit a "proposal," usually similar to a pre-analysis plan, though with more emphasis on the existing literature and discussion of the project's conceptual or theoretical contributions. Referees review this proposal, and the editor may engage in some back-and-forth with the submitting author, similar to the revise-and-resubmit process in a regular article submission. If the editor decides that the study is valuable and meets the journal's quality bar, it is awarded an "in-principle acceptance," similar to a conditional acceptance. The authors then analyze their data, write up results, and submit the full paper for stage two review.

In the second stage, the full paper is submitted with results, interpretation, and any extensions beyond the original plan (which are acceptable as long as they are clearly delineated). The key idea behind pre-results review is that the journal has committed to publishing the paper as long as the results are presented credibly, the interpretation is reasonable, and there were no major data problems along the way (which would drop the paper below the journal's standard for publication). For instance, if you tried to carry out a study in a country that then experienced a civil war or a natural disaster and you were unable to collect most data, the editor might decide the in-principle acceptance was no longer valid. But if the endline data looks to be of sufficiently high quality and the interpretation given to results is sensible, then the journal is committing to publishing the final paper even if the results are not statistically significant, challenge conventional wisdom, are surprising, or do not seem to "hang together" with a single unambiguous theoretical interpretation.

Virtually no social science journals used pre-results review in 2013, but the numbers have risen quickly with approximately 100 journals accepting "registered

reports" in 2018 and 277 journals today (Hardwicke and Ioannidis 2018).[9] In economics, the most prominent example of pre-results review is in the *Journal of Development Economics* (under editors Andrew Foster and Dean Karlan and with support from BITSS), starting in May 2018 (Foster et al. 2019). The *JDE* was a natural venue for a pilot given the already widespread use of pre-analysis plans in development economics, and to my knowledge, it is the first economics journal to adopt pre-results review as a standard article submission format.[10] As of January 2021, roughly two-and-a-half years in, the *JDE* had received 90 submissions for pre-results review, with a rising rate over time. Of these, 18 have received in-principle acceptance and three have been accepted in stage 2 and are now forthcoming in the journal, while the others are either undergoing stage-2 review, still assembling their data, carrying out analysis, or writing up the paper.

As part of the pre-results review adoption process at the *Journal of Development Economics*, a BITSS staff member (Aleksandar Bogdanoski) carried out phone interviews with 12 submitting authors to gain a qualitative sense of how pre-results review is being perceived (Foster et al. 2019). The interviews indicate that, despite being slightly different than regular articles, the refereeing process for pre-results review submissions went smoothly overall and no major red flags were raised, in part perhaps because detailed explanatory materials had been prepared in advance (for authors and referees), as well as a suggested template for the proposals. The most commonly cited benefit, by far, was that writing the proposal for peer review forced authors to think through their research design more carefully, and feedback from the referees at that early stage helped to further improve it. Another pattern was that junior scholars—particularly those who are going on the job market or up for tenure promotion—appreciate the ability to obtain an in-principle acceptance for a project that has not yet been completed (BITSS 2020).

Open questions remain regarding how pre-results review might work in subfields other than development economics. One other economics journal, *Experimental Economics*, has launched a pre-results pilot. The recent rise of alternative article formats in economics, inspired by the short format approach pioneered by *American Economic Review: Insights*, may facilitate acceptance of other novel approaches like pre-results review.

A distinct and lighter-touch change in journal policy are editorial statements to make a particular issue salient. In 2015, the editors of eight leading health economics journals issued an editorial statement emphasizing the importance of publishing

[9] This data is as of February 21, 2021. Up to date information on the adoption of pre-results review and registered reports can be found at https://www.cos.io/rr.

[10] The earliest pre-results analysis and pre-results review in economics (to my knowledge) is Neumark (2001), based on a one-off attempt to implement a pre-results review process at one economics journal in the 1990s. In 1996, there were heated minimum wage debates between Card and Krueger and Neumark and colleagues. According to my Berkeley colleague David Levine, who was the editor of *Industrial Relations* at the time, the late Alan Krueger had the idea in 1996 for various participants in the minimum-wage literature to pre-specify their analysis before the next federal wage increase, and as editor of *Industrial Relations*, Levine would commit to publish results (Levine 2001). Levine believes the idea originated from Danny Kahnemann, who in the 1980s and 1990s developed what he called "adversarial collaboration" with colleagues who disagreed with him, and with whom he worked together to design lab experiments and write articles. Christensen, Freese, and Miguel (2019) contains a more detailed discussion.

null results. They sent letters to referees reminding them to judge papers based on design and quality, not on whether the results are statistically significant.

Such a statement may seem like a small step, but it clearly encouraged a shift of norms. Blanco-Perez and Brodeur (2020) compare the share of published null results during 2014–2018 in the eight health economics journals to two applied microeconomics journals with no similar editorial statement. Figure 4 presents their data from the pre-period, the period when the editorial statement was implemented, and the post-period. The light gray line captures the share of papers presenting statistically significant results (at the 5 percent level) in the control journals and the dark gray line captures this proportion in the journals affected by the editorial statement. Pre-statement, roughly 50 percent of articles have null results in both the control and treatment journals, but there is a sharp rise in the publication of null results after the 2015 statement, with the share of null results increasing by 18 percentage points. This is due to some combination of changes in both editor and referee behavior; Blanco-Perez and Brodeur (2020) do not find meaningful changes in the characteristics of the papers submitted by authors to these journals over the study period.

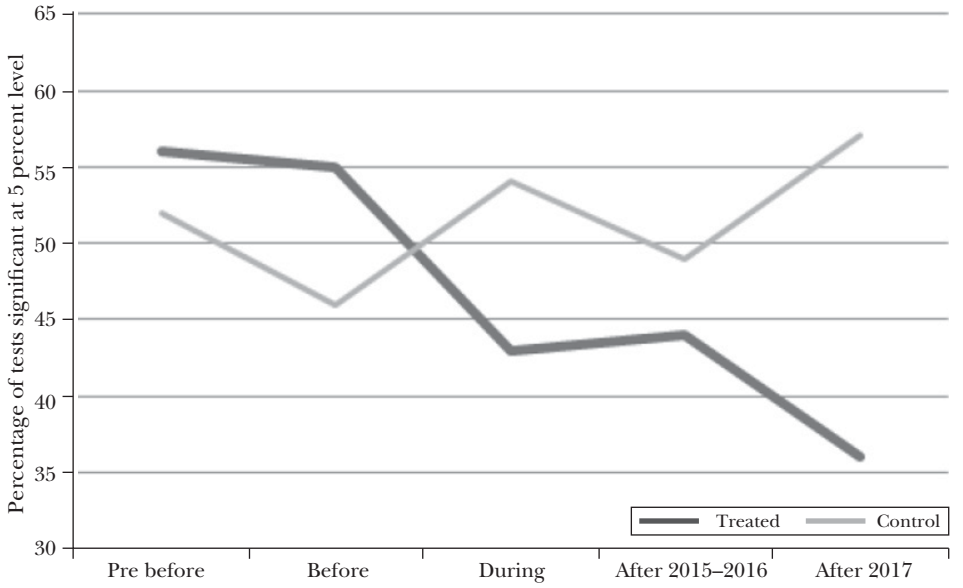
Of course, one can raise questions concerning the possibility of other changes that were occurring in these journals, or in the field of health economics, over time. Yet this evidence suggests that even simple and low-cost actions by editors might help promote changes in norms, even for something as deeply engrained as the bias in favor of publishing significant results. It seems worth considering similar editorial statements (with associated referee reminder letters) by other economics journals on the subject of null results, and perhaps on other open-science issues as well.

Looking Forward

The past two decades have seen rapid changes in policies and practices to promote open science in economics. Policies that were largely foreign to the discipline of economics when I was in graduate school—sharing of data and code, study registration and pre-analysis plans, and conflict-of-interest disclosure statements—are now routine parts of economists' workflows. Opening up the research process in economics promises to make our research more credible while also potentially promoting a more inclusive scholarly community. However, while some of the underlying problems of publication bias, specification search, and tendentious reporting may have receded, they have not yet gone away (Andrews and Kasy 2019; Brodeur, Cook, and Heyes 2020). In this essay, I have already mentioned some promising areas to enhance research transparency in economics. Here, I mention a few more.

In the area of pre-analysis plans, Laitin et al.'s (2020) plan to Report All Results Efficiently (RARE) proposes to make it standard practice for authors to post all results related to their pre-analysis plans on public study registries, in a so-called "pre-analysis plans report," even if those finding never make their way into a published paper (a proposal related to some ideas developed in parallel in Duflo et al. 2020 and Abrams, Libgober, and List 2020). This step would allow searches of study registries to yield far more complete evidence on work that has been carried

Figure 4
Journal Editorial Statements and the Publication of Significant Results



Source: From Blanco-Perez and Brodeur (2020, Figure 3).
Note: Treated journals include *Journal of Health Economics*, *European Journal of Health Economics*, *Health Economics*, *Health Economics Review*, and *International Journal of Health Economics and Management*. Control journals include *Journal of Public Economics* and *Labour Economics*. Percentage of tests significant at the 5 percent level by categories. "Pre Before the editorial" category includes papers that were published one year before the category "Before." "Before the editorial" category includes papers that were submitted and published before the statement on negative findings. "During the editorial" category includes papers that were submitted before the statement on negative findings, but published after. The "After the editorial" categories include papers submitted and published (respectively in 2015–2016 and 2017) after the statement on negative findings. Reproduced with permission from Abel Brodeur.

out on a certain topic to date, leading to improved meta-analysis as well as more informed choices for scholars launching new projects.

Another set of steps would seek to integrate preregistration approaches into some non-experimental research. For example, it might be possible to preregister studies of observational data in a way where it is possible to verify that the pre-analysis plan truly preceded the data analysis (for a discussion of this issue in medical research, see Dal-Ré et al. 2014). One can imagine a preregistration approach for studies that will be conducted after a particular event has occurred (such as an election or data release) or more generally before scholars have been granted access to restricted data (Burlig 2018; Christensen, Freese, and Miguel 2019). Ofosu and Posner (2020b) find that roughly 4 percent of pre-analysis plans that they reviewed were for observational data: in fact, among some studies discussed earlier, both Blanco-Perez and Brodeur (2020) about changes in journal editorial policies and Christensen, Dafoe et al. (2019) about impact of data-sharing were preregistered observational studies. The path to realistically utilizing

preregistration for a substantial share of observational nonprospective studies is uncertain, but remains a critical direction for future debate and innovation.

A cluster of other work is actively enriching preregistration in various ways, including by studies that compare effects of treatments with expert forecasts (Della Vigna and Pope 2018; Della Vigna, Pope, and Vivalt 2019), preregistering plans for split-sample analysis (Fafchamps and Labonne 2016; Anderson and Magruder 2017); or using a pre-analysis plan to guide the application of machine learning tools (Ludwig, Mullainathan, and Spiess 2019).

New ideas are also emerging about how to make reproducibility work better in economics: Lars Vilhuber (the Data Editor for the American Economic Association) is leading an effort in collaboration with the Berkeley Initiative for Transparency in the Social Sciences with the aim of Accelerating Computational Reproducibility in Economics (ACRE https://www.socialsciencereproduction.org/). The goal is a crowd-sourced platform to assemble and organize replication activities (which are often carried out today as graduate course assignments) in a systematic way, so that it can become possible to move away from black-or-white takes on whether a finding "replicates," and to illuminate the nuances involved in verifying empirical results. There is also a concrete proposal for how to bring more research transparency tools into public policy, termed Open Policy Analysis (Hoces de la Guardia, Grant, and Miguel 2018), which involves taking a specific policy analysis (say, Congressional Budget Office analysis of the effects of the minimum wage) and then fully specifying how the result was reached in an open-source online document that any member of the public can access.

Even as these open science tools expand in scope and influence, I think more work will also need to be done to change the culture and the mindset of the economics research community. In my opinion, economists should encourage ourselves, our colleagues, and our students to work on important problems without worrying so much about whether the results turn out to be immediately exciting: after all, if scholars are collecting good data and applying thoughtful methods while working on an important problem, even null results are meaningful. We should stress that all research conducted in this way contributes to the broader social goal of generating facts and learning about the world.

■ *I am grateful to Sarah Stillman and Simon Zhu for excellent research assistance. This article would not have been possible without my earlier collaborations with Aleks Bogdanoski, Kate Casey, Garret Christensen, Josh Cohen, Allan Dafoe, Andrew Foster, Jeremy Freese, Rachel Glennerster, Sean Grant, Fernando Hoces de la Guardia, Katie Hoeberling, Dean Karlan, David Laitin, Temina Madon, Don Moore, Betsy Levy Paluck, Andrew Rose and others, much of which is cited here. I also benefited from conversations with other colleagues at the Berkeley Initiative for Transparency in the Social Sciences, including Carson Christiano, Guillaume Kroll, Kelsey Mulcahy, Jen Sturdy, and Alex Wais. Gordon Hanson, Enrico Moretti, Timothy Taylor, and Heidi Williams provided useful suggestions. Some of the arguments here also feature in my 2019 NBER Summer Institute Methods Lecture: https://www.nber.org/econometrics_minicourse_2019/. All errors are my own.*

References

Abrams, Eliot, Jonathan Libgober, and John List. 2020. "Research Registries: Facts, Myths, and Possible Improvements." Artefactual Field Experiments Working Paper 00703.

Anderson, Michael L., and Jeremy Magruder. 2017. "Split-Sample Strategies for Avoiding False Discoveries." NBER Working Paper 23544.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Begg, Colin, Midred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, et al. 1996. "Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement." *Journal of the American Medical Association* 276 (8): 637–39.

Berkeley Initiative for Transparency in the Social Sciences (BITSS). (2020) "Exit Interviews with Authors of Stage 1 Proposals at the *Journal of Development Economics*." Unpublished data (accessed June 4, 2020).

Bernanke, Ben S. 2004. "Editorial Statement." *American Economic Review* 94 (1): 404.

Blanco-Perez, Cristina, and Abel Brodeur. 2020. "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* 130 (629): 1226–47.

Bogdanoski, Aleksandar, and Sarah Stillman. 2021. "Transparency and Openness Promotion (TOP) Factor Scores of Leading Economics Journals." Harvard Dataverse. https://doi.org/10.7910/DVN/JINXLP. (accessed February 8, 2021).

Brodeur, A., N. Cook, and A. Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." Unpublished.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Broockman, David, Joshua Kalla, and Peter M. Aranow. 2015. "Irregularities in LaCour (2014)." https://osf.io/preprints/metaarxiv/qy2se/.

Burlig, Fiona. 2018. "Improving Transparency in Observational Social Science Research: A Pre-analysis Plan Approach." *Economic Letters* 168: 56–60.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-analysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–1812.

Chang, Andrew C., and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'." Federal Reserve Board Finance and Economics Discussion Paper 2015-083.

Christensen, Garret, Allan Dafoe, Edward Miguel, Don A. Moore, Andrew K. Rose. 2019. "A Study of the Impact of Data Sharing on Article Citations Using Journal Policies as a Natural Experiment." *PLoS ONE* 14 (12).

Christensen, Garret, Jeremy Freese and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science.* Oakland, CA: University of California Press.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Christensen, Garret, Zenan Wang, Elizabeth Levy Paluck, Nicholas Swanson, David Birke, Edward Miguel, and Rebecca Littman. 2019. "Open Science Practices Are on the Rise: The State of Social Science (3S) Survey." *MetArXiv.* doi:10.31222/osf.io/5rksu.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Dal-Ré, Rafael, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. "Making Prospective Registration of Observational Research a Reality." *Science Translational Medicine* 6 (224).

De Angelis, Catherine D., Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzen, Christine Laine, et al. 2004. "Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors." *New England Journal of Medicine* 351 (12): 1250–51.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–29.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76 (4): 587–603.

Duflo, Esther, Abhijit Banerjee, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics." NBER Working Paper 26993.

Fafchamps, Marcel, and Julien Labonne. 2016. "Using Split Samples to Improve Inference about Causal Effects." NBER Working Paper 21842.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127 (3): 1057–1106.

Foster, Andrew, Dean Karlan, Edward Miguel and Aleksandar Bogdanoski. 2019. "Pre-results Review at the Journal of Development Economics: Lessons Learned So Far." *Development Impact*, July 15. https://blogs.worldbank.org/impactevaluations/pre-results-review-journal-development-economics-lessons-learned-so-far.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

Gerber, Alan, and Neil Malhotra. 2008a. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3: 313–26.

Gerber, Alan S., and Neil Malhotra. 2008b. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research*. 37 (1): 3–30.

Gertler, Paul, Sebastian Galiani, and Mauricio Romero. 2018. "How to Make Replication the Norm." *Nature* 554: 417–19.

Hardwicke, Tom E., and John P.A. Ioannidis. 2018. "Mapping the Universe of Registered Reports." *Nature Human Behaviour* 2: 793–96.

Hill, Ryan, Carolyn Stein, and Heidi Williams. 2020. "Internalizing Externalities: Designing Effective Data Policies." *AEA Papers and Proceedings* 110: 49–54.

Hoces de la Guardia, Fernando, Sean Grant, and Edward Miguel. 2018. "A Framework for Open Policy Analysis." *MetaArXiv*. doi:10.31222/osf.io/jnyqh.

Innovations for Poverty Action (IPA). (2020) "Operational metrics of the Research Transparency, Data Governance and Ethics Team," Unpublished data. (accessed June 2, 2020).

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8).

Kaplan Robert M., and Veronica L. Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time." *PLoS ONE* 10 (8): e0132382

Laitin, David, Edward Miguel, Ala Alrabahbah, Aleksandar Bogdanoski, Sean Grant, Katherine Hoeberling, Cecilia Mo, Don Moore, Simine Vazire, Jeremy Weinstein, Scott Williamson. 2020. "Unlocking the File Drawer: A RARE Proposal to Normalize Complete Reporting." Unpublished.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73 (1): 31–43.

Levine, David I. 2001. "Editor's introduction to 'The Unemployment Effects of Minimum Wages: Evidence from a Prespecified Research Design'." *Industrial Relations* 40 (2): 161–62.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2019. "Augmenting Pre-Analysis Plans with Machine Learning." Unpublished.

Mathieu, Sylvain, Isabelle Boutron, David Moher, Douglas G. Altman, and Philippe Ravaud. 2009. "Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials." *Journal of the American Medical Association* 302 (9): 977–84.

Merton, Robert K. 1942. "A Note on Science and Democracy." *Journal of Legal and Political Sociology* 1: 115–126.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K.M. Esterling, A. Gerber, R. Glennerster et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Neumark, David. 2001. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design the Employment Effects of MinimumWages." *Industrial Relations* 40 (1): 121–44.

Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25.

Ofosu, George K., and Daniel N. Posner. 2020a. "Do Pre-analysis Plans Hamper Publication?" *AEA Papers and Proceedings* 110: 70–74.

Ofosu, George K., and Daniel N. Posner. Forthcoming. "Pre-analysis Plans: An Early Stocktaking." *Perspectives on Politics*.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. "Sharing Detailed Research Data Is

Associated with Increased Citation Rate." *PloS One* 2 (3): e308.

Piwowar, Heather A., and Todd J. Vision. 2013. "Data Reuse and the Open Data Citation Advantage." *PeerJ* 1: e175.

Simonsohn, U., L.D. Nelson, and J.P. Simmons. 2014. "P-curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534–47.

Swanson ⓔ, Nicholas, Garret Christensen ⓔ, Rebecca Littman ⓔ, David Birke ⓔ, Edward Miguel ⓔ, Elizabeth Levy Paluck ⓔ, and Zenan Wang ⓔ. 2020. "Research Transparency Is on the Rise in Economics." *AEA Papers and Proceedings* 110: 61–65.

Turner, Erick. H., Annette M. Matthews, Eftihia Linardatos, Robert A. Tell, and Robert Rosenthal. 2008. "Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy." *New England Journal of Medicine* 358 (3): 252–60.

Vilhuber, Lars. 2020. "AEA Data and Code Availability Policy." *AEA Papers and Proceedings* 110: 776–78.

Vilhuber, Lars, James Turrito, and Keesler Welch. 2020. "Report by the AEA Data Editor." *AEA Papers and Proceedings* 110: 764–75.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 215–242

Why Is Growth in Developing Countries So Hard to Measure?

Noam Angrist, Pinelopi Koujianou Goldberg, and Dean Jolliffe

GDP and growth estimates in developing countries are often perceived to be fraught with problems and potentially manipulated. For example, when Nigeria's government changed the way it calculated GDP in 2014, the country's official GDP grew 89 percent overnight, raising concerns that the statistics were being manipulated for political purposes (as reported in *The Economist* 2014). This GDP revision updated the base year from 1990 to 2010, reflecting the economy's changing structure and giving greater weight to growing industries like mobile technology. In China, GDP growth estimates have been routinely called into question since the mid-2000s, and recent studies estimate that official Chinese statistics overstated average annual growth by 1.8 percentage points between 2010 and 2016 (Chen et al. 2019). In India, Subramanian (2019) estimates that changes

■ *Noam Angrist is a Fellow, Oxford University, Oxford, United Kingdom, Consultant, World Bank, Washington, DC, and Co-Founder, Young 1ove, Gaborone, Botswana. Pinelopi Koujianou Goldberg is Elihu Professor of Economics, Yale University, New Haven, Connecticut. She is also a Distinguished Fellow, Centre for Economic Policy Research, London, United Kingdom, Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, Executive Committee Member of the Bureau for Research and Economic Analysis of Development, and Non-resident Fellow at the Peterson Institute of International Economics. Dean Jolliffe is a Senior Advisor in the Development Data Group, World Bank, Washington, DC. He is also a Research Fellow, Institute for the Study of Labor (IZA), a Fellow, Global Labor Organization, and adjunct faculty with the School of Advanced International Studies. Their email addresses are noam.angrist@bsg.ox.ac.uk, penny.goldberg@yale.edu, and djolliffe@worldbank.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.215.

in data sources and methodology in 2011 led to an overestimation of annual growth by 2.5 percent between 2011 and 2012 and 2016 and 2017. Concerns are not confined to growth overestimation. Kerner, Jerven, and Beatty (2017) find that some lower-income countries underreport economic growth to maintain foreign assistance.

Is there robust evidence that growth is systematically mismeasured or measured less reliably in developing countries, or do such concerns reflect overgeneralizations based on a small number of widely publicized but nonrepresentative examples? More broadly, what specific challenges do developing countries face in measuring the growth of their economies? This article seeks to address these questions while offering thoughts on how growth measurement in developing countries can be improved.

There are plausible reasons why growth measurement could be more challenging for developing countries. Developing countries have lower statistical capacity, are often associated with weaker institutions and governance, have large informal sectors that are inherently hard to measure, and tend to be more reliant on agriculture. Volatile growth is harder to measure, and growth is more volatile in countries where agriculture constitutes a large part of the economy; this is especially true for rain-fed agriculture, which is highly correlated with low GDP per capita. Of course, advanced economies also face challenges: looking at US data, Aruoba et al. (2016) show that expenditure-side and income-side GDP estimates, though highly correlated, lead to different growth estimates. Likewise, Deaton (2005, Table 2) compares the average difference between GDP estimates based on national accounts and income estimates based on household surveys across countries, showing that the difference is *smallest* for countries in sub-Saharan Africa—though the coefficient of variation is also *greatest* for countries in sub-Saharan Africa, implying that changes over time are potentially more heavily influenced by measurement error in that region.

In this article, we first investigate the reliability of growth measurement across countries by comparing several data sources. We begin with a brief overview of GDP measurement and a discussion of the measurement challenges faced by all countries. We then triangulate and compare growth estimates based on several different data sources and methods: national accounts, household surveys, and satellite data on night-time light sensors and on vegetation mappings. While each source measures a different concept—so would not be expected to yield identical growth estimates—we interpret a tight concordance between different estimates as a sign of growth estimate reliability. We find that—contrary to common perceptions—there is no compelling evidence that growth is on average measured less well in developing countries. However, we find consistently higher dispersion in growth data for developing countries, which lends support to the view that perceptions about growth (mis)measurement may be due to higher levels of classical measurement error or the existence of a few problematic outliers.

We then turn to several measurement challenges specific to developing countries: limited statistical capacity, the use of outdated data and methods, large agricultural sectors, large informal economies, and limited price data. Using a newly constructed indicator of statistical integrity based on novel IMF audit data, we

do not find compelling evidence that statistical integrity is a first-order issue in most developing countries. We conclude by identifying concrete steps to improve growth measurement in developing countries, including strengthening statistical capacity and supplementing traditional growth measurement approaches with information from innovative data sources. For example, satellite-based vegetation data can measure activities by smallholder farmers that are less likely to be captured in GDP estimates, and several other new data sources offer scope to complement the standard methods. Overall, developing countries (especially low-income countries) perform better than expected at estimating output and growth given the constraints they face, but there is ample room for improvement.

A Brief History of National Income and Growth Measurement

While the notion of measuring economic growth has existed for centuries, today's commonly used methods are typically credited to the work of Simon Kuznets and Richard Stone. In the 1930s, the Great Depression created a desire to measure the severity of the crisis and any progress toward ending it (Kuznets 1934). In a powerful example of economic research informing policy, Kuznets reported on his work to the US Congress, and by 1942, the US government began publishing estimates of gross national product (GNP), in part to aid in war planning efforts. Around the same time, the United Nations (UN) recognized the value of measuring economic progress using methods that were consistent over time and comparable across countries. Stone helped the UN Committee on National Income Statistics develop a framework for a System of National Accounts (SNA) (Stone 1947a), and in 1953 the UN Statistical Commission released SNA guidelines that were applicable for most of the world, including developing or lower-income countries (Stone 1953). Both Kuznets and Stone would eventually receive the Nobel Prize for their work in developing and refining national growth accounting methods: Kuznets in 1971 (just the third Nobel Prize in economics ever awarded) and Stone in 1984 (https://www.nobelprize.org/prizes/economic-sciences/1971/kuznets/facts/; https://www.nobelprize.org/prizes/economic-sciences/1984/stone/facts/).

Since the original 1953 guidelines on the System of National Accounts, there have been a series of revisions to improve the quality of the measures and address measurement error, overseen by the Inter-Secretariat Working Group on National Accounts (ISWGNA)—a body comprising members from the International Monetary Fund (IMF), the European Union, the Organization for Economic Co-operation and Development, the UN, and the World Bank.[1] For example, following the most recent update to the SNA guidelines in 2008, the ISWGNA developed an *Implementation Programme for the System of National Accounts 2008 and Supporting Statistics* to assist

[1] For a history of these revisions, see Figure A1 in the online Appendix available with this article at the *JEP* website, or the UN Statistics Division website at https://unstats.un.org/unsd/nationalaccount/hsna.asp.

countries in building the statistical and institutional capacity needed to successfully transition to the new guidelines.

In addition to helping establish the System of National Accounts, Stone wrote seminal papers in the 1940s on measurement error in estimating national income. This early literature leveraged the variations in national income estimates from different measurement approaches (that is, expenditure-side and income-side) to assess and address measurement error (Stone, Champernowne, and Meade 1942). This approach is also the basis of recent literature, including Aruoba et al. (2016). Economists since Kuznets have long been familiar with the basic conceptual criticisms of GDP: that it fails to capture important aspects of well-being like leisure, health, and environmental protection, for example, or that it omits information about the distribution of income (Sen 1985; Nussbaum 1987; Stiglitz, Sen, and Fitoussi 2009).[2]

Despite concerns over measurement and interpretation, for decades nearly all countries worldwide have reported GDP and used the measure as a critical factor for short- and long-term policymaking.

Are Growth Estimates Less Reliable in Developing Countries?

There is no single, well-defined metric to assess the reliability of a country's national income and growth statistics. The most common approach, similar to Stone, Champernowne, and Meade (1942), is to compare growth estimates obtained using different data sources and approaches to examine whether the estimates coincide or are correlated. In this article, we explore three main conceptual constructs for the estimation of economic growth and make comparisons among them to assess the reliability of growth estimates.

The central measure we examine is GDP per capita, estimated based on System of National Accounts standards and usually produced by each country's National Statistical Office. As taught in introductory economics classes, GDP can be viewed as the sum of personal consumption, investment (including change in inventories), government expenditures, and net exports. Alternatively, it can be viewed as the sum of personal income, tax revenues on production and imports, and corporate tax revenues (including undistributed corporate profits).

We compare this standard measure to two alternative approaches. First, we consider household surveys of budgets, income, expenditure, or consumption. The

[2]Sen (1985) and Nussbaum (1988) argue that well-being is linked to the capability of an individual to live a life the person has reason to value. This is interpreted as being able to live a healthy life and participate in society without shame. This capabilities approach to measuring well-being underpins the United Nations Human Development Index. The Stiglitz, Sen, and Fitoussi (2009) critique of GDP is twofold. First, GDP fails to account for the within-country distribution of income. Second, some actions increase GDP but reduce well-being (like traffic jams leading to higher fuel consumption and a reduction in wel-lbeing), and similarly, some activities contribute to well-being but do not increase GDP (like unpaid household labor).

common method for this approach is to extract per capita household consumption or per capita household income, and then to compare growth rates of these measures with measures of personal income, personal consumption, or GDP per capita growth based on the System of National Accounts. A substantial share of household survey data is collected through large-scale efforts supported by the World Bank, such as the Living Standards Measurement Study. There are many reasons survey-based and SNA-based measures will differ. For instance, SNA protocols for income-side measures place relatively less emphasis on capturing informal economic activities, such as subsistence farming or so-called shadow activities such as the production of illegal drugs. Because household surveys in lower-income countries typically focus on asking people questions about what they have consumed (rather than what they have earned), they are more likely to capture such activities. Another difference is that SNA-based protocols place greater emphasis on larger transactions relative to smaller transactions, which have little impact on total income measures; in fact, Deaton (2005) documents that SNA training instructions directly specify that greater effort should be directed at larger transactions. In contrast, household budget and living standards surveys tend to include regular smaller transactions with greater probability than (often irregular) larger ones like weddings and funerals (Deaton and Zaidi 2002).

Next, we consider an approach that has only become possible in recent years: using satellite data for economic analysis (discussed in this journal by Donaldson and Storeygard 2016). Night-time lights have received particular attention, especially the Defense Meteorological Program (DMSP) Operational Linescan System (DMSP-OLS). Luminosity can serve as a proxy for economic activity, and night lights provide frequent, relatively cheap, and globally available data (Chen and Nordhaus 2011; Henderson, Storeygard, and Weil 2012; Pinkovskiy and Sala-i-Martin 2016).[3] Like other measurement approaches, night lights are imperfect. Zhou et al. (2015), for example, argue that limitations in the sensor of these lights create saturation problems in central urban areas, potentially hampering their ability to predict variation in economic activity in rich, high-density areas. By contrast, Gibson et al. (2021) argue that DMSP-OLS light data are poor predictors of economic activity in low-density, rural areas. An additional data source that potentially can be harnessed for growth measurement is satellite-based vegetation indices, estimated using reflectance from plants. A Normalized Difference Vegetation Index (NDVI) is estimated

[3]Night-time lights data are publicly available in an easy-to-use format from the National Oceanic and Atmosphere Administration (NOAA) website from 1992–2013. The site provides several data series. One frequently used night-time lights series is from the Defense Meteorological Satellite Program–Operational Linescan System (DMSP-OLS). This data source is cleaned to capture luminosity separate from the effects of cloud coverage, fires, aurora, and ephemeral light (Elvidge et al. 2009). Newer sources of night lights, such as the Visible Infrared Imaging Radiometer Suite (VIIRS), have also emerged; however, this data source is less regularly cleaned and is accessible for only a few years.

by satellite detection of reflectance from plants in specific portions of the visible and infrared spectra.[4]

Of course, one would not expect these various data sources to yield identical growth estimates. National accounts, household surveys, and satellite data were each designed for different reasons and serve different purposes. Nevertheless, we would expect the growth rates they generate to be correlated. Accordingly, in the following sections we examine correlations, long-run trajectories, and some key differences across these approaches. In the context of these comparisons, we examine whether growth-estimate reliability varies by country income grouping. In instances where such comparisons exist from earlier studies, we update them to more recent years and extend them to more countries.

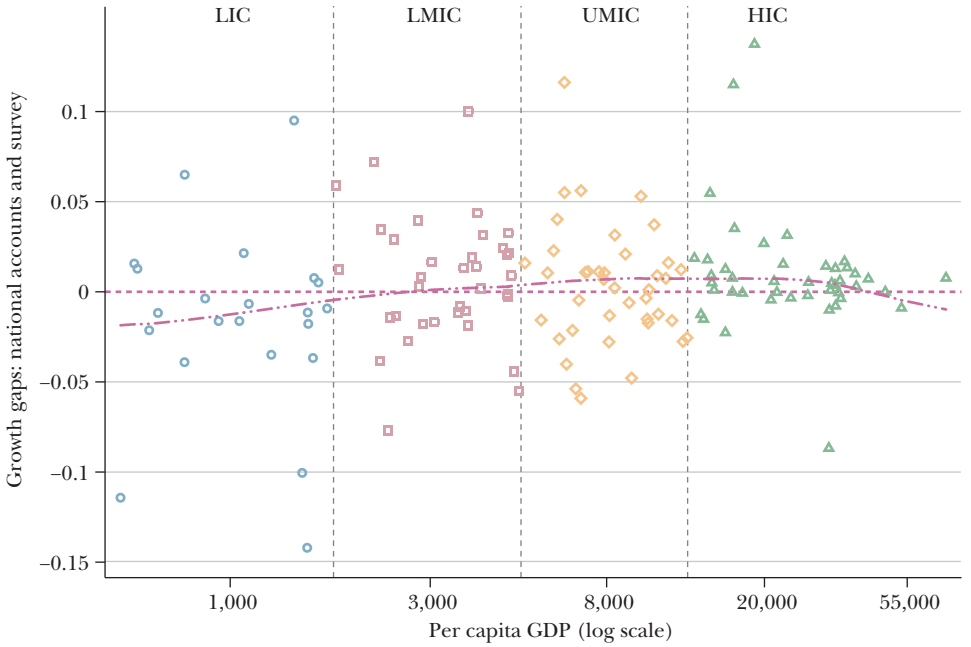
Growth Estimates from National Accounts and Household Surveys

It is well established that there are significant gaps between national accounts estimates of GDP or personal consumption and household survey estimates of income or consumption (Deaton 2005; Ravallion 2003). Prydz, Jolliffe, and Serajuddin (2020) updated this earlier analysis by examining the ratio in levels of GDP (and household final consumption expenditure) to income (and consumption) from a series of more recent household surveys, finding that middle-income—not low-income—countries have the weakest relationship between national accounts and survey measures. A potential explanation for this finding is that middle-income countries often have fast growth, which could decrease survey response rates (as households become richer, the opportunity cost of their time increases) and produce a downward bias in survey-based growth estimates. In addition, a more rapidly changing economic structure might increase discrepancies in income measurement if, for example, national accounts do not adjust the weights of industries that have become increasingly important over time (as was the case in Nigeria's 2014 GDP rebasing, for example). Broadly speaking, in the literature regarding GDP level estimates, there is an unresolved debate regarding the reliability of national accounts data by country income grouping, with conclusions varying among the leading studies.

Rather than examining levels, which have been examined in earlier papers and which are expected to differ across data sources given that different data measure different concepts, we focus our comparisons on growth rates. Growth-rate comparisons are, in principle, subject to the same caveats regarding differences in concepts measured, but we expect these caveats to be less consequential given that the focus on growth rates controls for the impact of time-invariant differences across measures. Also, annual growth rates often receive the most coverage and attention

[4]We use a data series for 89 economies where over 25 percent of employment is in agriculture from 2000 to 2018. We include measures for total Normalized Difference Vegetation Index per year per country as well as the maximum versus minimum NDVI in a given year and country. Based on definitions from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), we also disaggregate the NDVI by smallholder farms, which are often part of the informal economy versus large-scale commercial agricultural land, which usually is captured in national accounts.

Figure 1
Gaps in Average Growth between National Accounts and Survey Estimates



Source: Author calculations using data from the World Bank.
Note: A positive gap indicates GDP growth rates are higher than household surveys and vice-versa. Growth estimates are averaged over the time period 1992–2012 and are not weighted by the population of each country. Each income category is labeled: LIC = Low-income country; LMIC = Lower-middle-income country; UMIC = Upper-middle-income country; HIC = High-income country.

in international policy dialogues. However, measures of year-on-year growth are often volatile, and their variations are potentially due to noise. To minimize the impact of noise on our comparisons, we average annual growth rates over the time period 1992–2012 for each country.

Figure 1 provides a comparison of average growth rates based on estimated GDP per capita from national accounts and estimated per capita income or consumption from household surveys. For the household survey measure, we extract data from the World Bank's PovcalNet which provides a mix of per capita household consumption and income measures, depending on what is available (at http://iresearch.worldbank.org/PovcalNet/home.aspx).[5] For each country, we estimate

[5]The majority of these data files are based on integrated household surveys such as those in the Living Standards Measurement Study. The consumption aggregate is a broad measure, which includes consumption of food and nonfood items, with food consumption including food purchased from the market, home-produced food, and payment-in-kind. Nonfood consumption typically includes the total value of

the average growth rate over 1992–2012. The figure plots the gap between the two by income grouping. The most notable feature is the dispersion by income category, which is visibly highest for low-income countries and lowest for high-income countries. While the gap between different growth estimates is not significantly higher for developing countries on average, it is very large for select low-income countries.

As with the literature on GDP levels, the comparison of growth rates based on national accounts and survey data does not offer clear-cut conclusions. Our results seem consistent with the view that growth measurement may be most problematic in low-income countries, though as noted earlier, this view is supported more by the high dispersion of growth estimate gaps in low-income countries than by the size of the average gap.

GDP, Household Surveys, and Night-Time Lights Data

Next, we add into the analysis average growth rates based on satellite-based night light data by country from 1992 to 2012. While night light data have limitations and are an imperfect proxy for economic activity, they have two notable advantages: they are not biased by potential non-response, as household surveys are, and they are not easily manipulated or frequently adjusted, as national accounts data might be. Figure 2 plots a smoothed nonparametric regression of the growth rate based on each measure on log GDP per capita.

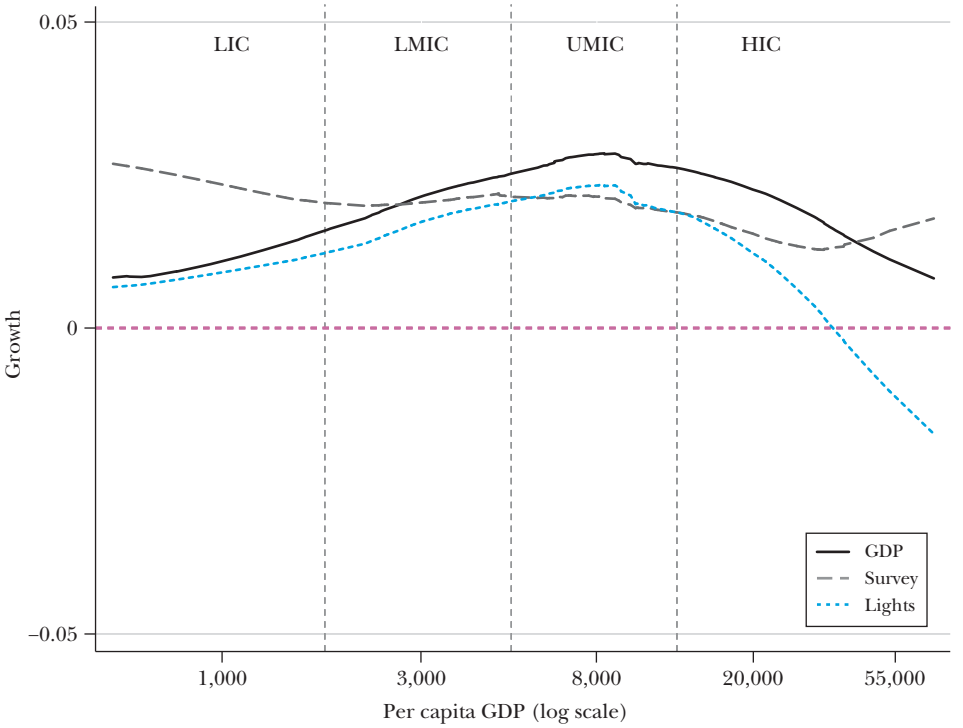
We observe a few patterns in the data. First, the GDP line shows the growth rate of per capita GDP over the range of countries based on national accounts data, which shows that middle-income countries grow more quickly than either high-income or low-income countries.

Second, growth rate estimates based on survey data are lower than estimates based on national accounts data for all categories *except* low-income countries, where survey estimates are higher: survey-based growth is on average 2.6 percent while national accounts-based growth is slightly less than 1 percent. One reason for this pattern might be that survey estimates capture more informal economic activity, which comprises a large share of the economies of low-income countries and which national accounts estimates may be less suited to measuring.

Third, light growth tracks GDP growth closely in all income categories except high-income countries. This suggests that lights might be useful in triangulating accurate GDP estimates in developing countries but that the relationship might be less clear for high-income countries. This pattern has several potential explanations: growth measurement could be less reliable in high-income countries; urban saturation in high-income countries might dampen light growth estimates; or the relationship between lights and economic growth (as measured by either national accounts or household surveys) could be non-monotonic by income level. For

small nonfood items plus the use-value of durable goods. For high-income countries, the majority of the PovcalNet data comes originally from either Eurostat's Statistics on Income and Living Conditions or the Luxembourg Income Study, which creates an income vector that is harmonized across countries in their archives.

Figure 2
Average Growth across Measures: GDP, Survey, and Lights



Source: Author calculations using data from the World Bank and NOAA.
Note: We run a LOWESS smoothed nonparametric regression of growth rates by income level on log GDP per capita terms. Each income category is labeled: LIC = Low-income country; LMIC = Lower-middle-income country; UMIC = Upper-middle-income country; HIC = High-income country. The categorization of countries is based on the current World Bank classification. For details of the calculations, including average growth rates for each measure and standard deviations across countries and over time, see the online Appendix.

example, some high-income countries have tried to reduce light pollution, in which case light would have a negative rather than positive association with income.

Fourth, while the gap between survey and national accounts is largest in low-income countries, for lights the gap is smallest in low-income countries. Hence, whether one considers growth measurement to be more or less reliable in developing countries may depend on which alternative measure one trusts most: lights or surveys.

However, if we examine variation across countries, we find a more consistent pattern, with high variation among low- and middle-income countries in the gaps between national accounts and surveys as well as lights. In Table 1, cross-country variation in national accounts growth estimates ranges from 2.2 to 3.8 percent

Table 1
Average Growth across Measures—GDP, Survey, and Lights

	Low income	*Lower middle income*	*Upper-middle income*	*High income*
Growth GDP				
Mean	0.009	0.025	0.028	0.023
Across country SD	0.038	0.022	0.031	0.018
Within country SD	0.052	0.044	0.050	0.030
Growth survey				
Mean	0.026	0.016	0.026	0.013
Across country SD	0.047	0.032	0.027	0.031
Within country SD	0.055	0.062	0.084	0.059
Growth lights				
Mean	0.009	0.017	0.025	0.011
Across country SD	0.028	0.032	0.047	0.023
Within country SD	0.183	0.152	0.159	0.196
Observations	25	38	39	40

Note: Growth rates are calculated as the log first difference. We average growth rates per country from 1992 to 2012 for each measure to account for year-to-year noise and variation. We than average country average growth rates by income category. Averages are not weighted by the population of each country. "Across Country SD" refers to the standard deviation of growth rates across countries, averaged by income category. "Within Country SD" refers to the standard deviation of growth rates *over time* within a country, averaged by income category.

in low- and middle-income countries, respectively, relative to 1.8 percent in high-income countries. Similarly, we observe higher cross-country variation in developing countries for survey and lights data. This points to a potential "black sheep" explanation: while discrepancies in growth estimates are not systematically worse in developing countries on average, there are a few countries for which such discrepancies are particularly large, and these cases may be responsible for the perception that growth measurement in developing countries is unreliable.

We observe a similar pattern for within-country variation of GDP growth estimates based on System of National Accounts data over time, ranging from 4.4 percent to 5.2 percent in low- and middle-income countries, compared to 3 percent in high-income countries. Again, this evidence suggests variation and volatility might play an important role in perceived reliability of growth estimates in developing countries.

Finally, we examine the role of limited data availability in some countries. We find that within-country correlations of survey and national accounts over time are higher at higher-income levels, varying from 0.16 to 0.33. Table A1 in the online Appendix, available with this article at the *JEP* website, shows within-country, year-to-year correlations between measures. However, this pattern virtually disappears when restricting the sample to countries with survey data for more than three time periods. Hence, it seems that the lower year-to-year correlations in

low-income countries are driven by limited data. For example, Rwanda has only three survey data points, meaning that growth rates can only be estimated at two points in time, and any correlation in estimates between survey and national accounts over time is derived from the single difference in growth from 2005 to 2010. As another example, Figure A2 in the online Appendix shows only five household survey data points in Tanzania between 1992 and 2012, relative to over 20 in Indonesia.

Summary

While some statistics suggest less reliable growth measurement in developing countries, the cumulative evidence is mixed. Previous work exploring correlations in GDP levels has not found evidence that low-income countries underperform higher-income countries in measurement. Similarly, we do not find systematic evidence based on night lights data that growth is measured less well on average or manipulated in developing countries. Light estimates in low-income countries follow a similar trajectory as national accounts estimates, and if anything, they track each other more closely than in high-income countries. In general, different comparisons lead to different conclusions. These results reinforce the value of supplementing national accounts estimates with survey-based measurement and of utilizing alternative sources of income estimates, such as satellite data, as we discuss in the paper's final section.

However, a consistent finding across all comparisons is that cross-country dispersion in growth estimates is substantially higher in developing countries, suggesting a possible role for a few outliers to generate the perception that all developing countries' growth estimates cannot be trusted.

Finally, we note that differences in average growth rates across the three different measurement approaches appear small—typically around 1.5 percentage points or less. While gaps of this magnitude may be considered large for high-income countries, where annual growth rates have recently been in the 3–4 percent range, they appear small for many fast-growing developing countries. We conclude that even though growth estimates may be imprecise, they are likely trustworthy within a margin of error of about 1.5 percentage points. Considering the uncertainty around such estimates, this error margin does not seem grave. It also suggests that paying excessive attention to potentially noisy year-on-year growth estimates seems unnecessary. At a minimum, year-on-year growth estimates should be accompanied by confidence intervals, which should be given as much attention as the estimates themselves.

Measurement Challenges in Developing Countries

While measurement challenges exist for both developed and developing countries, in this section we turn to the specific challenges that developing countries face in estimating growth.

Low Statistical Capacity and Lack of Independence of Statistical Authorities

The term "developing" signifies vulnerabilities and resource constraints that affect many areas, including data collection and production of statistics (Carletto, Jolliffe, and Banerjee 2015; Devarajan 2013; Jerven and Johnston 2015). Many developing countries use old data, outdated methods, and unreliable statistics due to lack of funding, inadequate resources for data collection, management and dissemination, and absence of coordination among relevant agencies and stakeholders. Statistical capacity constraints are particularly relevant in Africa (Devarajan 2013). As of early 2021, only about one-third of sub-Saharan African countries use the most recent System of National Accounts standards from 2008, while most of the rest use the 1993 standards.

Changing from one vintage of the System of National Accounts to another, or infrequent updates to growth accounting methods, can lead to substantial GDP movements, which in turn may contribute to the perception of unreliable or manipulated growth measurement in developing countries. For example, Ghana's adoption of the 1993 SNA system in 2010 led to a 62 percent upward revision of GDP (Devarajan 2013), and Ghana has since adopted the 2008 SNA. A similar revision in Malawi led to a 32 percent upward GDP revision. Likewise, failing to regularly update the base year for GDP estimation, which determines the weights reflecting the relative importance of different sectors, can create discrete breaks in a country's GDP series.[6] In addition to the aforementioned case in Nigeria, other examples include Senegal's 2014 rebasing (from 1999), which increased GDP by 29 percent, and Zimbabwe's 2012 rebasing (from 2009), which increased GDP by 20 percent. For a systematic view of countries' statistical capacity, in 2004 the World Bank developed the Statistical Capacity Indicator (SCI). Scores range from 0 (no statistical capacity) to 100 (adequate statistical capacity), with an overall score as well as scores in three sub-categories: Source Data, Methodology, and Periodicity.[7] The SCI's source data are collected mostly for low- and middle-income countries.

The average Statistical Capacity Indicator score for a low-income country is about 60, which is similar to the average regional score for the Sub-Saharan Africa and Middle East & North Africa regions. Lower- and upper-middle income countries have an average score of about 70, which is similar to the average regional score for Latin America & Caribbean and South Asia. Several low- and middle-income

[6]The US Department of Commerce's Bureau of Economic Analysis (BEA) introduced chain-weighting specifically to overcome this problem of discrete changes in GDP trends from occasional updates to a fixed base year (Steindel 1995).

[7]The source data for the Statistical Capacity Indicators refers to surveys for agriculture, health, poverty, the population census, and vital registration systems. The Methodology sub-category considers the following components: balance of payments manuals, consumer price index base year, external debt reporting status, government finance accounting, import and export price indices, industrial production price indices, national accounts base year, national immunization coverage, special data dissemination standards, and UNESCO reporting. The Periodicity sub-category refers to regular data for multiple categories including education, health, sanitation, and gender equality as well as GDP. The SCI Dashboard provides information on the time series of SCI, so that one can track countries' progress towards statistical capacity (http://datatopics.worldbank.org/statisticalcapacity/SCIdashboard.aspx).

countries, such as South Africa and India, score well on statistical performance, indicating that lower income is not synonymous with bad data.[8] The World Bank recently released the Statistical Performance Indicators (SPI), an update and re-conceptualization of the SCI.

As an alternative way to assess countries' statistical capacity and data quality, the IMF recently released a rich new dataset with information gathered in the process of compiling growth statistics (Berry et al. 2018). In contrast to the Statistical Capacity Indicator scores, which includes multiple statistics not directly linked to growth (such as education statistics reported to UNESCO), the IMF data focuses exclusively on data behind the System of National Accounts and also includes high-income countries. We observe some notable trends. First, the average SNA vintage is consistently older in low-income countries, aligned more closely to the 1993 guidelines than the more recent 2008 vintage commonly used in high-income countries. In addition, the GDP base year is older in low- and middle-income countries, which, as noted, increases the likelihood that national accounts will fail to reflect important changes in a country's economic structure (while also increasing the likelihood of large and potentially contested GDP expansions when the base year is ultimately updated). Second, while "availability of annual GDP" is similar across income categories, "availability of quarterly GDP" estimates varies substantially by income level, ranging from 38 percent for low-income countries to 91 percent for high-income countries. Third, the share of countries that independently compile GDP using different approaches (for example, based on expenditure and on production), which can enhance the reliability and quality of national accounts statistics, also varies by income level: 12 percent for low-income countries, 30 percent for lower-middle-income countries, 40 percent for upper-middle income countries, and 76 percent for high-income countries. A variety of other indicators are available in the IMF data, including timely release of annual or quarterly GDP data and advance release calendars.[9]

Here, we present a novel database of indicators based on expert audits of national accounts called the Reports on the Observance of Standards and Codes, which is a large initiative carried out jointly by the World Bank and the IMF to monitor compliance with international standards for statistical systems (for details, see https://www.imf.org/en/Publications/rosc). These reports assess criteria of the IMF Data Quality Assessment Framework (DQAF) for 83 countries. A main advantage of this new database is that it identifies additional quality measures that go beyond a focus on GDP compilation practices: as one example, there is an indicator related to revision policy and practice, which are viewed by the IMF as central to data quality. Each indicator is assessed by IMF auditors based on four rankings:

[8] For details of these calculations, along with a map showing these patterns by country, see the online Appendix available with this paper at the *JEP* website, especially Figure A3 and Table A2.

[9] The summary statistics presented here are compiled and structured from text responses to periodic IMF surveys conducted with 189 countries globally. We average statistics by income category. For detailed tables, see the online Appendix available at the *JEP* website.

Table 2
Summary Statistics for Systems of National Accounts: Capacity, Quality, and Integrity

	Quality		*Capacity*		*Integrity*		
	Revision	*Monitoring and process*	*Data use*	*Resources*	*Statistical professional practice*	*No prior data access*	*Legal environment*
High income	0.92	1.00	1.00	0.88	1.00	1.00	0.96
Upper-middle income	0.96	0.96	1.00	0.71	0.96	0.83	0.92
Lower-middle income	0.95	1.00	0.75	0.65	0.95	0.95	0.85
Low income	0.80	0.90	0.60	0.30	1.00	0.90	1.00
East Asia & Pacific	0.86	1.00	1.00	0.71	1.00	1.00	1.00
Europe & Central Asia	0.97	0.97	1.00	0.80	0.97	0.97	0.97
Latin America & Caribbean	0.93	1.00	0.93	0.67	1.00	0.93	1.00
Middle East & North Africa	0.86	1.00	0.86	0.86	0.86	0.86	0.86
North America	1.00	1.00	1.00	1.00	1.00	1.00	1.00
South Asia	1.00	1.00	0.50	0.75	1.00	1.00	0.50
Sub-Saharan Africa	0.87	0.93	0.67	0.40	1.00	0.80	0.87

Note: This table summarizes novel data compiled by the World Bank and IMF and aligned to the United Nations Fundamental Principles of Official Statistics. IMF staff routinely conduct in depth audits with countries around the world including visits to National Statistics Offices and joint review of data sources and process documentation. We group a subset of the indicators arising from these audits displayed in the left-hand column of online Appendix Table B1, available at the *JEP* website, to three high-level categories: Quality (indicators 4.3 and 0.4); Capacity (indicators 5.1 and 0.2); and Integrity (indicators 1.1, 1.2, and 0.1). Table B1 in the online Appendix includes more background on each indicator.

observed, largely observed, largely not observed, or not observed. For our purposes, we code analysis as a dummy variable equal to one if the practice is observed or largely observed, and zero otherwise.

Table 2 breaks down seven indicators from this new database by income group and region. The first two columns show measures of quality: whether revisions and updates of GDP estimates follow a regular and transparent schedule and whether they are monitored and accompanied by explanatory notes. Low-income countries appear to have lower-quality statistics, which is consistent with the indicators of statistical capacity already presented. For example, 80 percent of low-income countries have revision policies and practices, compared to 92 percent and 96 percent in high- and middle-income countries, respectively.

The next two columns show measures of statistical capacity. We first examine human resources in national statistical offices. While 88 percent of national statistical offices in high-income countries are deemed to have enough human resources, this indicator falls to only 30 percent in low-income countries.

The final three columns seek to measure the potential for politically motivated data manipulation, referred to as data integrity. A surprising pattern in this category

is that the lowest scores are observed in middle-income countries. Only 83 percent of upper-middle income countries specify that there is no internal governmental access to statistics prior to their release. Moreover, only 85 percent of lower-middle income countries have a legal environment that clearly delineates responsibilities for the collection and processing of data, compared to 96 percent of high-income countries and 100 percent of low-income countries. These patterns suggest that manipulation may be more feasible where there exists a threshold level of statistical capacity and sophistication that can potentially be used to promote political agendas.

Overall, constraints on statistical capacity emerge as a major factor affecting the quality of implementing the System of National Accounts in low-income countries, while conditions for deliberate data manipulation are more likely to be observed in middle-income than low-income countries.

The Role of the Agricultural Sector

The agricultural sector contributes about 5 percent of total world economic production but represents a much larger share in most developing countries.[10] In Africa, agriculture is the largest sector and accounts for 15 percent of total GDP. In some developing countries, especially in Africa and South Asia, agriculture represents more than half of economic output (according to the World Development Indicators, https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?end=2019&start=1960&view=chart). In addition, agriculture's contribution to growth volatility is about three times greater than the service sector's contribution (Koren and Tenreyro 2007).

However, agricultural production is often poorly measured (Jerven and Johnston 2015; Carletto, Jolliffe, and Banerjee 2015). In many low- and middle-income countries, the quantity of crops harvested on cultivated land or the amount of land cultivated are estimated in part through self-reported farmholder surveys which suffer from significant levels of measurement error (Abay et al. 2019; Dillon et al. 2018; Gourley, Kilic, and Lobell 2019). For example, Carletto, Savastano, and Zezza (2013) show that self-reported plot sizes by the bottom decile of farmers (in terms of landholdings) are double what satellite measurements indicate. Similarly, the data used by Desiere and Jolliffe (2018) indicate that self-reported crop yields by the bottom quartile of farmers (in terms of landholdings) are about twice as large as actual yields.

Self-reports of the quantity and value of production are also fraught with measurement concerns. Many subsistence farmers sell relatively little of their crop output but are nonetheless frequently asked to report its value and quantity. When market transactions do inform responses, they are frequently based

[10] The CIA World Factbook estimates agriculture value added to be 6.4 percent while the World Bank's World Development Indicators estimates the value added of agriculture to global GDP to be about 4 percent (https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS). Figure A4 in the online Appendix available at the *JEP* website shows worldwide estimates.

in nonstandard units such as heaps, piles, buckets, or bags, which are often not comparable beyond a limited geographic area. For example, Capéau and Dercon (2006) note that a *tassa* (serving can) is commonly used to report market transactions in Ethiopia, but the unit of measurement is known to be significantly larger in northern Ethiopia.

For countries with large agricultural sectors, the reliability of estimated GDP growth depends on how well agricultural activities are accounted for in national accounts. As national accounts focus on measuring total output, the methodological approach places greater emphasis on accurately capturing large farms' production. Agricultural household surveys, by contrast, typically focus on understanding constraints to improving yields and profits for smallholder farms, which comprise a sizable share of agricultural activity. Lowder, Skoet, and Raney (2016) estimate that there are 570 million farms worldwide, over 87 percent of which are small (less than 2 hectares or about 5 acres) and family operated. Moreover, 95 percent of smallholder farms are in low- or middle-income countries.

We assess whether the high prevalence of smallholder farms in low- and middle-income countries reduces the reliability of GDP growth estimates. To do so, we compare GDP value-added agricultural growth in national accounts and agricultural growth proxied by a satellite-based vegetation index (see online Appendix available at the *JEP* website for a data description) from 2000 to 2018 across 87 countries.

Table A4 in the appendix shows regression results. We find a positive and statistically significant relationship of .317 for all farms, which suggests that the vegetation index is highly correlated with national accounts estimates of agricultural output. We next disaggregate the vegetation index by smallholder (column 2) and larger corporate agricultural growth (column 3), based on definitions from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS). While large corporate agricultural activity has an even stronger relationship with GDP estimates, reflected in a statistically significant coefficient of .388, smallholder growth has *no statistically significant relationship* with GDP estimates. Including country and time fixed effects (columns 4 and 5) leads to qualitatively similar results.

These results are visualized in Figure 3, which shows the positive relationship between agricultural output based on national accounts estimates and the vegetation index for all farms. This is driven by the positive relationship with large farm output, while there is a strikingly flat and slightly negative relationship with small farm output.

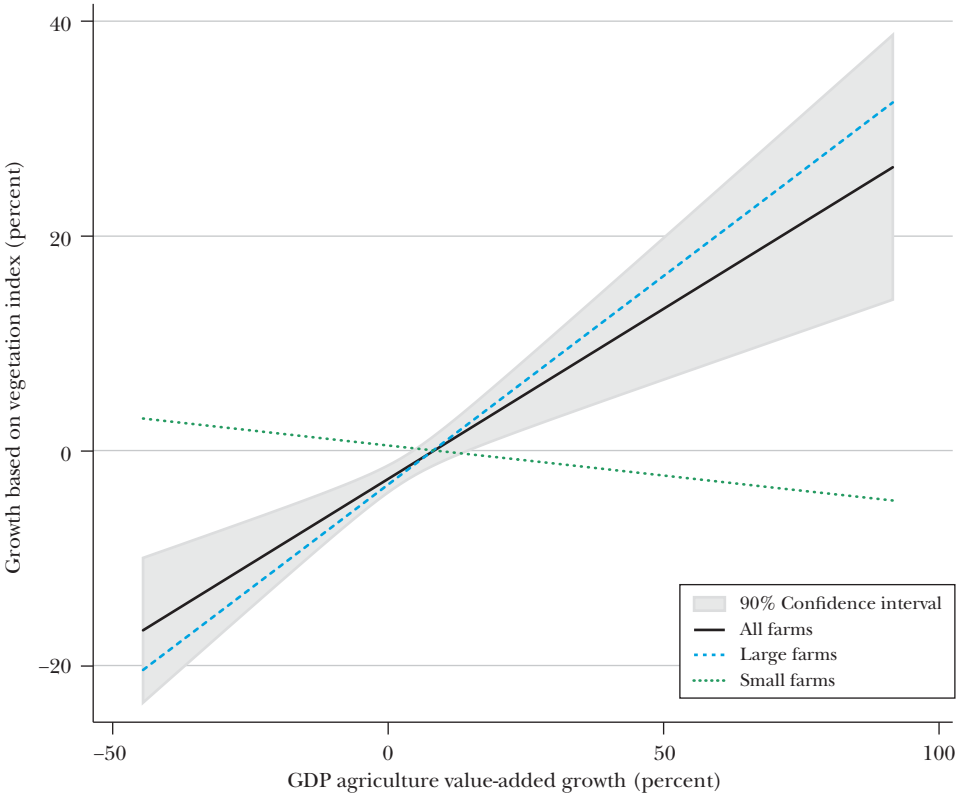
These results are consistent with the interpretation that smallholder agricultural activity is not well captured in official GDP as estimated in national accounts. This has substantial implications for the accuracy of growth measurement in developing economies, where smallholder farms are particularly important.

Informal or Shadow/Underground Economy

Developing countries are also characterized by a large informal sector, defined broadly as economic activity that is invisible to government, either because firms are

Figure 3

Measuring Agricultural Output: Comparing Growth Based on National Accounts and Satellite-Based Vegetation Growth



Source: Author calculations based on satellite images and data from Landstat8, and farm type classifications based on NASA definitions.

Note: Data for the agriculture vegetation index was produced from satellite imagery. The distinction between smallholder and larger corporate farms is based on definitions from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS). For a description of the data, see the online Appendix available at the *JEP* website. We run regressions using panel data from 2000 to 2018 across 87 countries in which the share of agricultural employment is above 25 percent.

not registered (and hence avoid taxes and regulations) or workers are not registered (and hence do not receive social protection). The concept of "informality" was born in Africa; in Ghana, "informal income opportunities" are common for individuals, and in Kenya, it is typical for enterprises to be informal (Charmes 2012).

The informal economy is also occasionally referred to as the "shadow" or "underground" economy. Illegal activities are typically not well captured in measures of GDP based on the System of National Accounts, though they are arguably important in some countries. For example, in Afghanistan, the drug industry is estimated

to comprise as much as one-third of GDP but is largely not accounted for in official growth statistics (Buddenberg and Byrd 2006). In contrast, farmer self-reports of poppy production in Afghanistan's national household survey are substantial and do not appear to suffer from significant nonresponse problems.[11]

The informal sector represents a major measurement challenge in developing countries—especially in sub-Saharan Africa—for the same reasons that agriculture is a challenge. This is in part because there is substantial overlap between agriculture and the informal economy. Based on data from household surveys in 69 countries, the International Labour Organization (2018) estimates that the informal economy represents 41 percent of GDP in sub-Saharan Africa, ranging from less than 30 percent in South Africa to 60 percent in Nigeria, Tanzania, and Zimbabwe.[12] Charmes (2012) reports that in the 2000s in sub-Saharan Africa, the informal sector (including the agricultural household sector) contributed nearly two-thirds of GDP, with the highest share in Niger (72.6 percent) and the lowest in Senegal (51.5 percent). Excluding agriculture, the informal sector represents approximately one-third of sub-Saharan Africa's GDP. In India, Charmes (2012) estimates that the informal sector comprises 54.2 percent of GDP (or 38.4 percent if agriculture is excluded). Using data from 158 countries from 1991 to 2015, Medina and Schneider (2018) estimate the average size of the "shadow" economy to be 31.9 percent of GDP, with the highest shares in Zimbabwe (60.6 percent) and Bolivia (62.3 percent). In sum, while specific estimates vary, existing work indicates that the share of the informal economy in low-income countries is substantial.

The contribution of informal enterprises to GDP can be measured in multiple ways, including surveys of establishments and households or by the residual difference between national expenditure and income statistics. However, since the contribution of informal labor employed in formal enterprises (as an intermediate input) is not included in GDP measurement of final output, this approach likely results in underestimates of the contribution of the informal sector to GDP. Moreover, it also likely results in underestimates of growth, as it is generally believed that informal employment in formal enterprises is growing in developing countries. This underestimation is more pronounced in countries with both large informal employment and a large number of formal enterprises, which tend to be middle-income economies.[13] Overall, these considerations suggest that the mismeasurement of the

[11] Buddenberg and Byrd (2006) provide several explanations for this, one of which is that there is a tradition of openness about discussing poppy production in part due to the legacy from when opium bazaars were common and out in the open. There is also the issue that the household interview is about crop production and not drug production, and that the poppy is just one crop of many that the farm households are asked about.

[12] For estimates of the informal economy worldwide, see Figure A5 in the online Appendix available at the *JEP* website.

[13] On the other hand, as Charmes (2012) points out, the way the informal sector is treated in the SNA-based measures of GDP may also lead to overestimates of its contribution to GDP because current measurement practice is premised on the assumption that the household sector can be assimilated into the informal sector. This assumption may be true in low-income countries characterized by subsistence

informal sector's contribution to growth may be a bigger issue in middle-income than low-income countries.

These measurement challenges are presumably biggest during policy changes that affect the formal and informal sectors differently. For example, in India real gross value-added growth for the informal sector is proxied by the Index of Industrial Production, which is mostly composed of formal sector firms. While this approach may work reasonably well during normal times, it likely overstated growth in the aftermath of India's demonetization and the Goods and Services Tax (GST)—both policy changes that have been shown to have disproportionately impacted the informal sector (Subramanian 2019; Chodorow-Reich et al. 2018).

Price Measurement

Price deflators are needed to obtain changes in real GDP, but prices are often poorly measured in developing countries. For example, a recent controversy in Rwanda regarding poverty measurement resulted from differences in inflation measurements: while the Consumer Price Index (CPI) suggested that Rwanda's inflation rate from 2011 to 2014 was 23 percent, the National Institute of Statistics in Rwanda (NISR) used a 4.7 percent inflation rate to calculate poverty rates. There were also substantial differences in inflation rates between urban and rural areas, which are largely not captured in the official price index (as reported by Wilson and Blood 2019). In India, Subramanian (2019) flags that the use of a manufacturing Wholesale Price Index as a proxy for producer prices of services in the mid-2010s, a time of sharply declining oil prices, could have led to gross value-added and real growth being overstated.

Some prominent data series on national income lack underlying data on price levels, particularly in developing countries. Young (2012) notes that in 2006 the UN National Accounts database providing GDP estimates in current and constant prices was missing more than half of all 1,410 observations across 47 sub-Saharan African countries from 1991 to 2004. Moreover, among 15 of the countries for which the complete time series are published, there was no constant price data. Similarly, Young (2012) notes that the purchasing power parity index in the Penn World Tables (PWT) version 6.1 provides real incomes for 45 sub-Saharan African countries, but 24 do not have a benchmark study of prices. In 2005, the World Bank's International Comparison Program (ICP) measured prices for 146 countries, for the first time including many previously-excluded developing countries. Accordingly, a substantial revision was conducted between PWT 6.1 and PWT 7.0 to include this new price data, resulting in large differences between countries in per capita income and larger growth estimates for many countries, especially in Africa (Young 2012).

The IMF has collected a dataset to assess statistical practices for price indexes in 193 economies along a variety of dimensions (Berry et al. 2019). The data show

agriculture and a small formal sector, but is less likely to hold in emerging economies with larger and growing formal sectors.

that while consumer price indexes are available in all economies, compliance with the international standard Classification of Individual Consumption According to Purpose (COICOP) system varies substantially by income category: there is 92 percent adoption in high-income countries, and between 57 and 75 percent adoption in low- and middle-income countries. While 87 percent of high-income countries have national expenditure coverage in their consumer price index, only 62 percent do so in low-income countries, with a substantial share (25 percent) of countries deriving price information from capital cities only.[14]

When looking at information on producer price indices in this IMF data, we find a steep gradient of data availability by income category. Availability of information on producer price indexes is 79 percent for high-income countries and only 41 percent for low-income countries. The timeliness of producer price index data follows a similar pattern, with monthly data available for 63 percent of high-income economies but only 12 percent of low-income economies. In terms of alignment to a recent classification system vintage, such as to the International Standard Industrial Classification of All Economic Activities (ISIC) Revision 4, 56 percent of high-income economies align with this vintage, compared to less than 9 percent of low-income economies. In addition, when developing a producer price index, the IMF recommends starting with the mining, manufacturing, and utilities sectors, and expanding coverage to more complex activities, such as services over time. We find that while around 60 percent of high-income countries include at least the mining, manufacturing, and utilities sector, only 16 percent of low-income economies do so. No low-income country includes any sectors beyond mining, manufacturing, and utilities.[15]

Finally, we examine the practice of "inflation targeting" which refers to the central banking policy aimed at achieving a specific annual rate of inflation. This practice is seen to be a strong proxy for the quality of national accounts systems as it provides a direct incentive for national statistics offices and government ministries to have accurate and timely price information (Carson, Enoch, and Dziobek 2002). We find that while 65 percent of high-income countries practice inflation targeting, only 12 percent of low-income countries do. This indicator can be viewed as a summary statistic for many of the more specific price indicators, as quality and timeliness of each of the specific price indicators makes this practice possible.

Altogether, we find strong evidence from the IMF data that price data in lower income countries is often lacking, out-of-date, or not aligned to best practice.

[14] For detailed data on price indexes by country and the Classification of Individual Consumption According to Purpose system, see Table A5 in the online Appendix available with this paper at the *JEP* website.

[15] For a table showing a more detailed list of price index practices compiled by the IMF across 193 economies by Berry et al. (2019) as well as a breakdown by high-income, upper-middle income, lower-middle income, and low-income countries, see the online Appendix available with the paper at the *JEP* website.

Do Measurement Challenges Explain Gaps between Alternative Growth Measures?

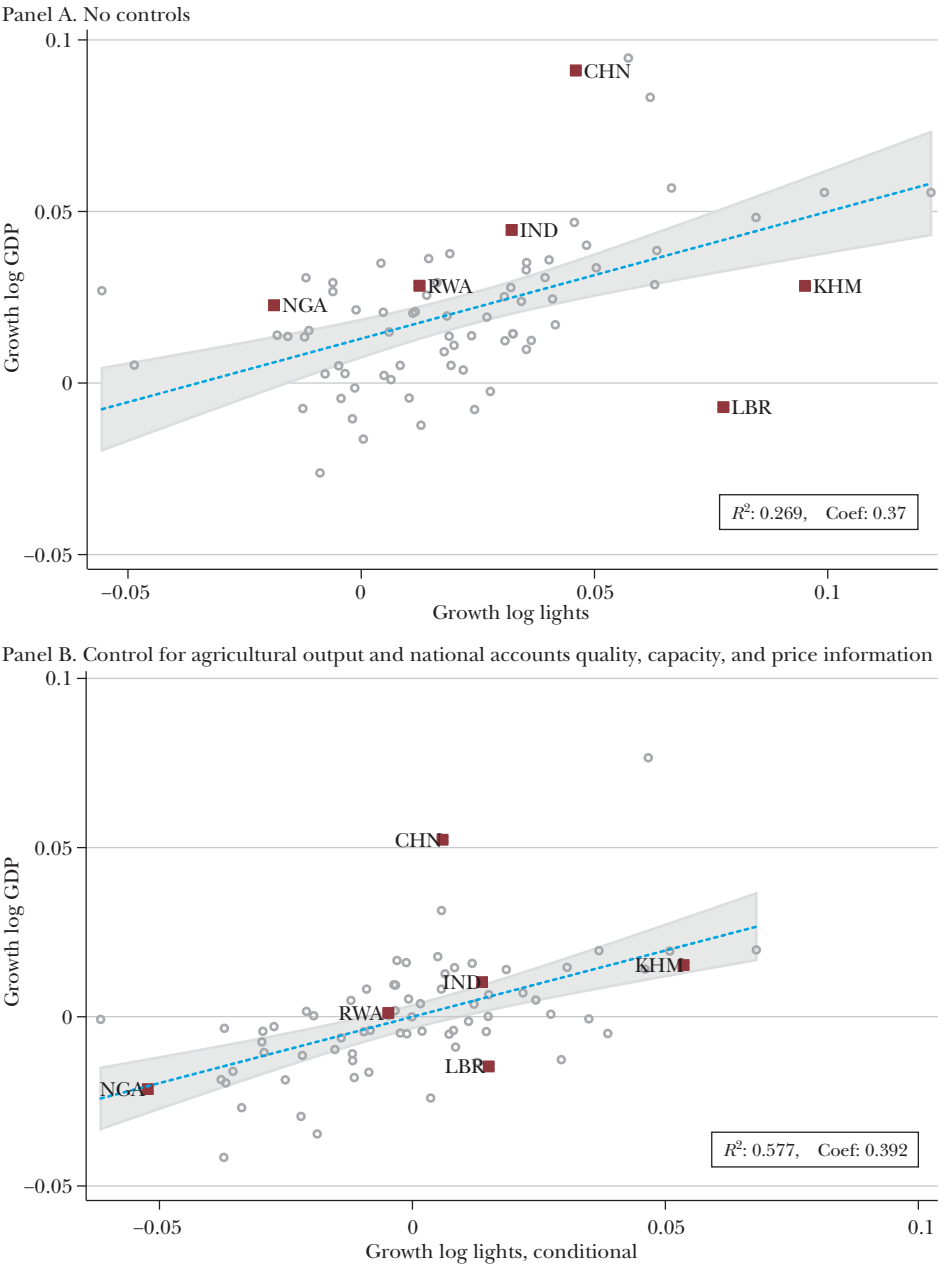
We now examine how each factor discussed in the previous subsections influences the reliability of growth measurement, as proxied by concordance among various growth measures. For example, when we compare GDP and lights data, the average elasticity between the growth estimates based on these two data sources is around 0.37. If a country's GDP is substantially higher than the best-fit line, this raises a flag that the country might be manipulating its GDP estimates; GDP can be manipulated to higher numbers for political purposes but satellite-based night-lights data cannot.

Figure 4 shows in Panel A the average elasticity between lights and GDP, illustrating deviations from the average elasticity for a select group of countries: China, India, Rwanda, Nigeria, Liberia, and Cambodia. Notably, the GDP growth estimates for China, India, Nigeria, and Rwanda—each of which have faced controversies regarding their statistics—lie *above* the line of best fit, which is consistent with the idea that these countries might be reporting higher growth relative to real economic activity for political purposes. However, this divergence could also be due to other factors; for example, using an inaccurate price index to calculate real GDP could inflate GDP relative to real economic activity. In the case of Cambodia and Liberia, which lie *below* the line of best fit, the divergence might be explained by the countries' large informal economies, which can be observed by night lights but are not fully accounted for in GDP estimates.

We examine whether controlling for factors that we suspect may be responsible for growth mismeasurement reduces the divergence from the average elasticity and increases the *R*-squared of the associated regression. In Figure 4, we focus on a subset of 74 countries which heavily rely on agriculture (defined as a share of employment in agriculture that is over 25 percent). The unconditioned correlation in Panel A between the log of growth in GDP and night lights suggests a series of countries have growth rates that differ substantially from what is predicted by lights data. Controlling for a series of other indicators, including the satellite-based vegetation index (which plausibly captures smallholder agricultural economic activity) as well as agricultural value-added in national accounts and price measurement practices (Panel B), results in a tighter concentration around the fitted line as revealed by the substantial increase in *R*-squared from 0.269 to 0.577. Several countries (for example, Cambodia, India, Liberia, Nigeria, and Rwanda) are no longer outliers. This suggests that the divergence observed in Panel A may have been driven by the presence of smallholder agriculture, the informal economy, and challenges in measuring price changes. Notably, China does not converge substantially, suggesting the plausibility of GDP data manipulation.

In the online Appendix, available with this paper at the *JEP* website, we conduct this exercise with 164 countries (see Figure A6), successively adding more control variables which helps further explain the difference between night-lights data and the SNA-based measures of GDP. For example, when we condition on our IMF data that is based on the Reports on the Observance of Standards and Codes, we no longer observe any outliers among the 60 countries for which we have data (Figures

Figure 4
Comparing GDP and Lights with and without Vegetation Index Controls



Source: Author calculations based on data from the World Bank, vegetation satellite data from Landstat8, as well as quality, capacity and price data from the IMF.
Note: Figure 4 includes average growth for 74 countries from 1992 to 2012 for lights and GDP. Panel A plots the bivariate correlation of the log growth of GDP and lights. Panel B conditions this relationship on the vegetation index, quality, capacity, price measurement practices, agricultural value-added in national accounts, and the share of GDP attributed to natural resources.

A7 in the online Appendix). This suggests that when the aforementioned challenges of measuring GDP are accounted for, the correlation between night-lights data and GDP is high. In short, the measurement challenges reviewed in this paper matter substantially and can help explain discrepancies in growth measurement..

How Can We Do Better?

What are some concrete steps that could improve growth measurement in developing countries? While some constraints may be political, such as policymakers who may not be interested in statistical practices that could make them look bad, good measurement can also shine a light on progress and reveal fruitful areas for policy action. Duly noting the political constraints, we now discuss a few areas for improvement.

Improve Statistical Capacity

Improving statistical capacity is an obvious and frequent recommendation, but also a challenging one. International efforts to support national statistics offices are often focused on one-off data collection activities with limited attention to building the skills and knowledge of national statisticians or to developing data systems. Collecting data is a relatively well-defined task with a clear end date that usually wraps up with a completion report. Investments to improve statistical capacity are much more difficult to monitor, less certain to succeed, time-consuming, and often lacking clear outcome deliverables.

Infrequent GDP rebasing is one specific problem facing many developing countries that would be feasible to address. Moreover, when countries do update their GDP base years, they often do not adequately explain or document the changes; the resulting GDP volatility contributes to perceptions of possible data manipulation. While the 1993 SNA guidelines state a preference for moving away from fixed base-year methodologies towards annual chain indices, they recognize that some countries with limited statistical capacities will need to continue following fixed-base year methods. For these countries, the base year should be updated annually and then estimates should be linked across base years to maintain comparability of trend data (IMF 1993). This approach keeps reference prices (and thereby implicit weights) current, while also smoothing out discrete GDP breaks.

Combine Traditional Data with Innovative Data Sources

An explosion of new and publicly available data sources has taken place over the last decade: web-scraping, Google searches, digital transactions, mobile phone metadata, social media usage, satellite data, and others. There are important examples of these sort of data outperforming traditional data sources: for example, Blumenstock, Cadamuro, and On (2015) use mobile phone metadata to estimate poverty and wealth, and Cavallo and Rigobon (in this journal, 2016) use web-scraped price data to estimate inflation. These new sources of data are illuminating and useful

but should be viewed as complements rather than substitutes for traditional data for several reasons.

First, national income accounting relies on a wide array of data sources including data collected by other government agencies for administrative purposes, national surveys, and censuses. Most of these data were collected for purposes other than national income accounting. For example, population census, agricultural census, industrial census, price surveys, household surveys, and labor force surveys were designed for other purposes (like reducing the harms of poverty, food insecurity, and unemployment). Even if replacing a traditional data source with a new one proved successful for the narrow purpose of estimating GDP, dropping or neglecting the traditional source would most likely damage the ability to fulfill its primary purpose.

Second, traditional data sources typically seek complete coverage of current populations, although they certainly face challenges in doing so, such as underrepresentation of informal settlements, slum inhabitants, and top-income earners. In contrast, while data from new sources can be massive in sample size and very timely, they are rarely representative of the population of a nation (for example, Blank and Lutz 2017).

Third, the joint use of traditional and newer data offers complementarities, as in the examples we include in this paper of supplementing GDP measurement with satellite-based data on night lights or vegetation yields. Another example is agricultural yield measurement: while traditional fieldwork is useful for obtaining estimates of average yield, satellite data can help improve estimates of yield variations (Lobell et al. 2020). Likewise, using satellite data to augment traditional sampling frames (Tollefson 2017) based on population censuses is another example of a useful hybrid approach. The modal frequency for population censuses is once every ten years; a common practice is to survey samples at annual or more frequent intervals within geographic areas, then use the decadal census-based population weights to extrapolate annual results for the country. Sampling frames based on population censuses are often inaccurate even when fresh, because of coverage problems (particularly in densely populated areas and informal settlements), and they become outdated over time. Cross-country analysis by the Bongaarts and Bulatao (2000) finds that population counts from censuses are off by 3 percent on average in the year the count was carried out, and that the five-year projections from the base year are off by 6 percent. Supplementing population frames with a combination of satellite-based estimates of housing structures and on-ground sampling of inhabitants per typical structure can provide more accurate estimates. More accurate population estimates would have a direct role on GDP per capita estimates and could also feed into future GDP measurements.

Monitor Performance, Identify Gaps, Offer Transparency

Just as countries collect data to monitor the performance of their policies and programs, collecting metadata on national data and statistical systems would also have value. As noted, in 2021 the World Bank released the Statistical Performance

Indicators (SPI) as an upgrade to the earlier Statistical Capacity Index. Although the goal to measure the capacity of national statistical systems is the same, the new SPI has expanded into new areas including data use, administrative data, geospatial data, data services, and data infrastructure. Continuing efforts to improve the quality of assessments of data systems can identify weak links and thereby target resources for improved measurement.

In addition, the IMF regularly collects detailed information from countries on their practices with regard to the System of National Accounts, including GDP revision policies, data access prior to public release, and GDP compilation and public release practices. Much of this data exists in open-response text form and is publicly available on the IMF website for over 140 countries. As noted, the IMF recently codified a subset of this information into easy-to-analyze datasets (Berry et al. 2018; Berry et al. 2019). The IMF also conducts SNA audits, with detailed reports available publicly online for 83 countries. In this paper, we collaborated with the IMF to codify information available in these audits to create a usable dataset for the first time. Efforts similar to this one, which harness the global reach and infrastructure of institutions such as the IMF and the World Bank, could substantially improve information on national data and statistical systems.

Our analysis, and others like it, clearly show that many countries are not following the latest guidelines and compliance is far from complete. Poor transparency, including lack of commitment to open and easily-accessible data, is just as critical to address. Making data available to the public requires investing staff time and skill for documentation (including codebooks, field manuals describing protocols, sample design, and metadata on coverage and response), de-identifying and preparing the data for safe dissemination, and other steps. This requires a culture of documenting and publicly disclosing the decisions made and methodologies used in GDP estimation. Just as "sunlight is the best disinfectant," transparency limits both the scope for and perception of political manipulation of data.

■ *We are grateful to Michael Stanger and Jim Tebrake of the IMF and to Hogeun Park of the World Bank for expert guidance and a compilation of some of the data used in this study as well as for many helpful discussions. We also thank JEP co-editors Heidi Williams and Gordon Hanson as well as Timothy Taylor for their guidance, and Josh Blumenstock, Paul Collier, Angus Deaton, Stefan Dercon, John Gibson, and David Weil for useful inputs and conversations. Rhea Gupta provided excellent research assistance and Greg Larson provided invaluable editorial support. We are grateful to the Economic Growth Center (EGC) at Yale for financial support. The views expressed here are those of the authors and should not be attributed to the World Bank.*

References

Abay, Kibrom A., Gashaw T. Abate, Christopher B. Barrett, and Tanguy Bernard. 2019. "Correlated Non-Classical Measuremgent Errors, 'Second Best' Policy Inference, and the Inverse Size-Productivity Relationship in Agriculture." *Journal of Development Economics* 139: 171–84.

Aruoba, S. Borağan, Francis X. Diebold, Jeremy Nalewaik, Frank Schorfheide, and Dongho Song. 2016. "Improving *GDP* Measurement: A Measurement-Error Perspective." *Journal of Econometrics* 191 (2): 384–97.

Berry, Francien, Brian Graf, Michael Stanger, and Mari Ylä-Jarkko. 2019. "Price Statistics Compilation in 196 Economies: The Relevance for Policy Analysis." International Monetary Fund Working Paper 19/163.

Berry, Francien, Massimiliano Iommi, Michael Stanger, and Louis Venter. 2018. "The Status of GDP Compilation Practices in 189 Economies and the Relevance for Policy Analysis." International Monetary Fund Working Paper 18/37.

Blank, Grant, and Christoph Lutz. 2017. "Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram." *American Behavioral Scientist* 61 (7): 741–56.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile phone Metadata." *Science* 350 (6264): 1073–76.

Bongaarts, John, and Rodolfo A. Bulatao, eds. 2000. *Beyond Six Billion: Forecasting the World's Population*. Washington, DC: National Academy Press.

Buddenberg, Doris, and William A. Byrd, eds. 2006. *Afghanistan's Drug Industry: Structure, Functioning. Dynamics, and Implications for Counter-Narcotics Policy*. Washington, DC: United Nations Office on Drugs and Crime.

Capéau, Bart, and Stefan Dercon. 2006. "Prices, Unit Values and Local Measurement Units in Rural Surveys: An Econometric Approach with an Application to Poverty Measurement in Ethiopia." *Journal of African Economies* 15 (2): 181–211.

Carletto, Calogero, Dean Jolliffe, and Raka Banerjee. 2015. "From Tragedy to Renaissance: Improving Agricultural Data for Better Policies." *The Journal of Development Studies* 51 (2): 133–48.

Carletto, Calogero, Sara Savastano, and Alberto Zezza. 2013. "Fact or Artifact: The Impact of Measurement Errors on the Farm Size–Productivity Relationship." *Journal of Development Economics* 103: 254–61.

Carson, Carol S., Charles Enoch, and Claudia Helene Dziobek, eds. 2002. *Statistical Implications of Inflation Targeting: Getting the Right Numbers and Getting the Numbers Right*. Washington, DC: International Monetary Fund.

Cavallo, Alberto, and Roberto Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives* 30 (2): 151–78.

Charmes, Jacques. 2012. "The Informal Economy Worldwide: Trends and Characteristics." *Margin: The Journal of Applied Economic Research* 6 (2): 103–32.

Chen, Wei, Xilu Chen, Chang-Tai Hsieh, and Zheng Song. 2019. "A Forensic Examination of China's National Accounts." NBER Working Paper 25754.

Chen, Xi, and William D. Nordhaus. 2011. "Using Luminosity Data as a Proxy for Economic Statistics." *Proceedings of the National Academy of Sciences* 108 (21): 8589–94.

Chodorow-Reich, Gabriel, Gita Gopinath, Prachi Mishra, and Abhinav Narayanan 2018, "Cash and the Economy: Evidence from India's Demonetization." NBER Working Paper 25370.

Deaton, Angus. 2005. "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)." *Review of Economics and Statistics* 87 (2): 395.

Deaton, Angus, and Salman Zaidi. 2002. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." LSMS Working Paper 135.

Desiere, Sam, and Dean Jolliffe. 2018. "Land Productivity and Plot Size: Is Measurement Error Driving the Inverse Relationship?" *Journal of Development Economics* 130: 84–98.

Devarajan, Shantayanan. 2013. "Africa's Statistical Tragedy." *Review of Income and Wealth* 59 (S): S9–15.

Dillon, Andrew, Sydney Gourlay, Kevin McGee, and Gbemisola Oseni. 2018. "Land Measurement Bias and Its Empirical Implications: Evidence from a Validation Exercise." *Economic Development and Cultural Change* 67 (3): 595–624.

Donaldson, Dave, and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in

Economics." *Journal of Economic Perspectives* 30 (4): 171–98.

The Economist. 2014. "How Nigeria's Economy Grew by 89% Overnight," *The Economist,* April 8. https://www.economist.com/the-economist-explains/2014/04/07/how-nigerias-economy-grew-by-89-overnight.

Elvidge, Christopher D., Edward H. Erwin, Kimberly E. Baugh, Daniel Ziskin, Benjamin T. Tuttle, Tilottama Ghosh, and Paul C. Sutton. 2009. "Overview of DMSP nighttime lights and future possibilities." In *2009 Joint Urban Remote Sensing Event,* 1–5. Piscataway, NJ: IEEE.

Gibson, John, Susan Olivia, Geua Boe-Gibson, and Chao Li. 2021. "Which Night Lights Data Should We Use in Economics, and Where?" *Journal of Development Economics* 149.

Gourlay, Sydney, Talip Kilic, and David B. Lobell. 2019. "A New Spin on an Old Debate: Errors in Farmer-Reported Production and Their Implications for Inverse Scale–Productivity Relationship in Uganda." *Journal of Development Economics* 141.

Henderson, J. Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

International Labor Organization. 2004–2016. "Proportion of Informal Employment in Total Employment by Sex and Sector (%)." *International Labor Organization* https://www.ilo.org/shinyapps/bulkexplorer25/?lang=en&segment=indicator&id=SDG_0831_SEX_ECO_RT_A. (accessed September 7, 2020).

International Labor Organization. 2018. *Women and Men in the Informal Economy: A Statistical Picture.* Geneva: International Labour Office.

International Monetary Fund. 1993. *System of National Accounts, 1993.* Washington, DC: International Monetary Fund.

International Monetary Fund. 2001–2019. "Reports on the Observance of Standards and Codes (ROSCs)." https://www.imf.org/en/Publications/rosc.

Jerven, Morten, and Deborah Johnston. 2015. "Statistical Tragedy in Africa? Evaluating the Data Base for African Economic Development." *The Journal of Development Studies* 51 (2): 111–15.

Kerner, Andrew, Morten Jerven, and Alison Beatty. 2017. "Does It Pay to Be Poor? Testing for Systematically Underreported GNI Estimates." *Review of International Organizations* 12: 1–38.

Koren, Miklós, and Silvana Tenreyro. 2007. "Volatility and Development." *Quarterly Journal of Economics* 122 (1): 243–87.

Kuznets, Simon. 1934. *National Income, 1929–1932.* Cambridge, MA: National Bureau of Economic Research.

Lobell, David B., George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. 2020. "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis." *American Journal of Agricultural Economics* 102 (1): 202–19.

Lowder, Sarah K., Jakob Skoet, and Terri Raney. 2016. "The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide." *World Development* 87: 16–29.

Medina, Leandro, and Friedrich Schneider. 2018. "Shadow Economies around the World: What Did We Learn Over the Last 20 Years?" IMF Working Paper 18/17.

National Centers for Environmental Information. 2012. "National Trends in Satellite Observed Lighting: 1992–2012". *National Oceanic and Atmospheric Administration* https://ngdc.noaa.gov/eog/dmsp/download_national_trend.html (accessed September 7, 2020).

Nussbaum, Martha. 1987. "Nature, Function, and Capability: Aristotle on Political Distribution." WIDER Working Paper 31.

Open Street Map. 2017. "ne_50m_admin_0_countries." https://puma.worldbank.org/layers/geonode%3Ane_50m_admin_0_countries (accessed September 7, 2020).

Pinkovskiy, Maxim, and Xavier Sala-i-Martin. 2016. "Lights, Camera…Income! Illuminating the National Accounts-Household Surveys Debate." *The Quarterly Journal of Economics* 131 (2): 579–631.

Prydz, Espen, Jolliffe, Dean and Umar Serajuddin. 2020. "Mind the Gap: Disparities in Assessments of Living Standards Using National Accounts and Surveys." Conference paper, IARIW-WB Conference on New Approaches to Defining and Measuring Poverty in a Growing World, Washington, DC.

Ravallion, Martin. 2003. "Measuring Aggregate Welfare in Developing Countries: How Well Do National Accounts and Surveys Agree?" *Review of Economics and Statistics* 85 (3): 645–52.

Sen, Amartya. 1985. *Commodities and Capabilities.* Amsterdam: North-Holland.

Steindel, Charles. 1995. "Chain-Weighting: The New Approach to Measuring GDP." *Current Issues in Economics and Finance* 1 (9): 1–6.

Stiglitz, Joseph, Amartya Sen, and Jean Fitoussi. 2009. *Report of the Commission on the Measurement of Economic Performance and Social Progress (CMEPSP).* https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf.

Stone, Richard. 1947. "Measurement of National Income and the Construction of Social Accounts." Report of the Subcommittee on National Income Statistics of the League of Nations Committee of Statistical Experts 7. Studies and Reports on Statistical Methods. Geneva, Switzerland: United Nations.

Stone, Richard, D. G. Champernowne, and J. E. Meade. 1942. "The Precision of National Income Estimates." *Review of Economic Studies* 9 (2): 111–25.

Subramanian, Arvind. 2019. "India's GDP Mis-Estimation: Likelihood, Magnitudes, Mechanisms, and Implications." CID Faculty Working Paper 354.

Tollefson, Jeff. 2017. "Satellite Images Reveal Gaps in Global Population Data." *Nature* 545 (7653): 141–42.

US Geological Survey. 2001–2018. "Landsat Normalized Difference Vegetation Index." https://www.usgs.gov/core-science-systems/nli/landsat/landsat-normalized-difference-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con (accessed September 7, 2020).

Wilson, Tom, and David Blood. 2019. "Rwanda: Where even Poverty Data Must Toe Kagame's Line," *Financial Times*, August 12. https://www.ft.com/content/683047ac-b857-11e9-96bd-8e884d3ea203.

World Bank. 1960–2019. "Agriculture, Forestry, and Fishing, Value Added." https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS (accessed September 7, 2020).

World Bank. 1960–2019. "World Development Indicators." https://data.worldbank.org/indicator/NY.GNP.PCAP.CD (accessed September 7, 2020).

World Bank. 1960–2019. "Poverty & Equity Data Portal." *PovcalNet* https://povertydata.worldbank.org/poverty/home/ (accessed September 7, 2020).

World Bank. 2004–2020. "Statistical Capacity Indicators." https://databank.worldbank.org/reports.aspx?source=Statistical-capacity-indicators (accessed September 7, 2020).

Young, Alwyn. 2012. "The African Growth Miracle." *Journal of Political Economy* 120 (4): 696–739.

Zhuo, Li, Jing Zheng, Xiaofan Zhang, Jun Li, and Lin Liu. 2015. "An Improved Method of Night-Time Light Saturation Reduction Based on EVI." *International Journal of Remote Sensing* 36 (16): 4114–30.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 243–256

Retrospectives
James Buchanan: Clubs and Alternative Welfare Economics

Alain Marciano

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact either Beatrice Cherrier, CNRS & CREST, ENSAE-Ecole Polytechnique (beatrice.cherrier@gmail.com) or Joseph Persky, University of Illinois at Chicago (jpersky@uic.edu).

Introduction

Club goods are characterized by non-rivalry in consumption, like pure public goods. Unlike pure public goods, however, club goods allow excludability in supply. In the case of a club good, a group of individuals both share the costs of provision of the good and limit its access, so that consuming the club good benefits only those who have paid a fee or a toll and are thus members of the club. This is the case of health or sport clubs, country clubs, as well as local public goods—such as swimming pools, museums, or libraries—but can also apply to groups of people using highways, the education system, hospitals, and the services of a police department or fire protection department.

James Buchanan (1965a) is considered the first to introduce this category of goods in economics. Buchanan's seminal article, "An Economic Theory of Club" (1965a), and its legacy have been widely discussed (in particular, see Sandler and

■ *Alain Marciano is University Professor of Economics, University of Montpellier, Montpellier, France. His email address is alain.marciano@umontpellier.fr.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.243.

Tschirhart 1980, 1997; Sandler 2013). However, the genesis of the idea of club goods has rarely been assessed from Buchanan's own perspective and work. This is what we do here, drawing on published and unpublished work and correspondence from the James M. Buchanan Papers archived at George Mason University.[1] This allows us to show the connection between Buchanan's work on the pricing of public goods and his criticism of Samuelson's welfare economics. Indeed, Buchanan invented the concept of club goods to support an alternative form of welfare economics, which could dispense with the use of a social welfare function. Clubs in the sense of Buchanan are thus foreign and incompatible with the traditional Samuelson-style public economics in which they are often used.[2]

Buchanan was convinced that social welfare functions—as used in Paul Samuelson's work that Buchanan targeted—were an arbitrary and unnecessary means to determine the individual contributions to the provision of a public good and to guarantee the desirability of an allocation of resources. Arbitrary, because the effect of social welfare function is to impose taxes independently from individual preferences and not needed because they ignore that individuals are willing to pay for the public goods they consume or, for that matter, the external effects they produce. Instead of using a social welfare function, Buchanan argued, individuals should pay a price based on their willingness to pay for public goods or external effects. Clubs were the institutional mechanism that would make the implementation of individualized prices for public goods or external effects possible.

This essay focuses on the development of Buchanan's views about clubs. We start with his beliefs about the pricing of public goods and the scope of welfare economics, and then proceed by explaining why and when clubs exist. We conclude with a discussion of the pricing mechanism Buchanan suggested for use in clubs.

Buchanan (and Wicksell) versus Samuelson on Welfare Economics of Public Goods

Fundamental principles of Buchanan's "fiscal philosophy"—drawing heavily on the work of the Swedish economist Knut Wicksell (1896, 1958; on Buchanan and Wicksell, see Marciano 2020)—shape his analysis of clubs: individualized prices for public goods, willingness to pay, and unanimity. To Buchanan, taxes are or should be viewed as prices—"taxes or contributions paid are exchanged for services rendered by the political unit" (1948, 38). Individuals buy public goods and services from the state as they buy private goods and services. Then, the principle of voluntarism held that individuals would pay these prices—even

[1] The James M. Buchanan papers are referred to as BP below and in the reference list [C0246, Special Collections Research Center, George Mason University Libraries].

[2] Buchanan's (1965a) article was unrelated to Charles Tiebout's "A Pure Theory of Local Expenditures" (1956). Buchanan (1957a, 1961b; Buchanan and Goetz 1972) disagreed with Tiebout's emphasis on mobility and spatial competition as a means to deal with free riding in public finance (for more details, see Boettke and Marciano 2017).

when it meant paying for the external costs their consumption creates—because they benefit from the goods and services they consume. Finally, the principle of unanimity held that the tax structure should be determined by asking all individuals how much they are ready to contribute.

Unanimity served two purposes. First, it sought a form of justice: in the words of Wicksell (1896, 114; 1958, 90), "if justice requires no more, it certainly requires no less. In the final analysis, unanimity and fully voluntary consent in the making of decisions provide the only certain and palpable guarantee against injustice in tax distribution." Unanimity was indeed "a guarantee against action designed to benefit or harm special classes unjustly" (Buchanan 1951, 176), a means to avoid the "tyranny of the majority" (Buchanan 1948, 11) and, beyond, to protect "the working and poorer" (Johnson 2010, 193). Second, as Buchanan (1951, 177) insisted, since "no one is worse off if some allocation of the required tax can be found which is acceptable to everyone," unanimity guaranteed that the "Paretian conditions for optimum welfare" (Buchanan 1951, 177) are satisfied without resorting to interpersonal comparisons of utility and *without using a social welfare function.*

These were the bases of Buchanan's welfare economics, the basis for his reaction to Samuelson's (1954) analysis of the optimal provision of public goods.

Samuelson (1954) famously established three conditions to be satisfied to reach a Pareto-optimal allocation of resources in an economy with private and pure public goods. First, there was the standard optimality condition for private goods. Second, a specific condition for public goods—the sum of the individual marginal rates of substitution between a public good and any private good should be equal to a unique marginal rate of transformation between those two goods. Third, a condition based on a social welfare function would determine how much each individual would pay—this condition encapsulated the "normative judgments concerning the relative ethical desirability of different configurations involving some individuals being on a higher level of indifference and some on a lower" (Samuelson 1954, 387). The second and third of Samuelson's conditions were the object of Buchanan's criticism.

Buchanan rejected social welfare functions because they meant that individuals would be coerced. A social welfare function rests on "ethical evaluations" or "value judgements" (Buchanan 1959, 133). It "is an explicit expression of a value criterion" (Buchanan 1959, 133). More precisely, the values embodied in the social welfare function are those of the observer who builds it—an economist, decision maker, or social welfare planner. Indeed, interpersonal comparisons of utility are unavoidable in building a social welfare function—a social welfare function necessarily "embod[ies] interpersonal norms" (Buchanan 1955a, 2) or "involv[es] interpersonal considerations" (Buchanan 1955a, 1). However, this social welfare function requires that preferences could be "read" by an external observer, who can then derive individuals' marginal rates of substitution between the public and a private good and aggregate them to determine the optimal amount of public good to produce. An external observer can only have access to individual preferences if it is assumed that the utility functions are revealed by previous behavior—if, as Buchanan wrote to

Samuelson, it is assumed that "individual ordinal preferences can be derived only by *revealed* choices" (Buchanan to Samuelson, February 25, 1955, BP; emphasis in original).

To Buchanan, however, individual preferences cannot be presumed before a choice is made. Preferences do not exist outside and independently from the choice itself. Buchanan wrote (1969, 72):

> [T]here is simply no means of determining, even indirectly, the value that they [individuals] place on the utility loss that might be avoided. In the classic example, how much would the housewife whose laundry is fouled give to have the smoke removed from the air? Until and unless she is actually confronted with this choice, any estimate must remain almost wholly arbitrary.

If no one can "read" preferences from behaviors, then the marginal rates of substitution supposedly "read" from the behavior of individuals, such as those used in the social welfare functions, are actually being imposed by the external observer. As Buchanan had written (1959, 133):

> Individual preferences, insofar as they enter the construction (and they need not do so) must be those *which appear to the observer* rather than those revealed by the behavior of the individuals themselves. In other words, even if the value judgments expressed in the function say that individual preferences are to count, these preferences must be those presumed by the observer rather than those revealed in behavior.

The external observer can be mistaken or biased—by imposing a value judgement—in the distribution of the costs of the public good. To avoid these pitfalls, Buchanan claimed that one should base the prices for public goods on individual's willingness to pay. This leads to the second set of criticism he raised against Samuelson.

Buchanan (1955a, 3) found Samuelson's abandonment of individualized prices for public goods and the use of a "unique marginal rate of substitution in production" to be problematic. This meant that Samuelson had abandoned price discrimination—"the relative 'prices' of collective and private goods are made uniform for all individuals." Removing price discrimination *for private goods* was "a step toward Pareto optimality" (Buchanan 1955a, 3), since consumers can adjust their consumption to equalize their marginal rate of substitution between two private goods to the relative price of these goods. However, removing price discrimination was a step away from Pareto optimality *when there are public goods*, because individuals cannot adjust their consumption of public goods. Indeed, the marginal rate of substitution between public and private goods is given. Thus, an individual whose marginal rate of substitution differs from the unique relative price will not be in an optimal situation, and an allocation of resources reached under Samuelson conditions—that is, based on this unique exchange ratio—could

hardly be optimal. Those who made little or no use of the public good could end up paying the same as those who extensively consume it. One could even envisage that one or a few individuals might bear the entire cost of provision of the public good (Buchanan to Samuelson, March 4, 1955, BP). Indeed: "It is true the summation of individual exchange ratios must be equal to the 'social' exchange ratio. But this does not allow the individual differences to be neglected, since there are many possible ways of adding up" (Buchanan 1955a, 5).

In Buchanan's view, prices for public goods should be individualized to allow an adjustment that was impossible to achieve through quantities. This was why Buchanan added an *individual* condition to satisfy a guarantee of an optimal provision of public goods. Buchanan's individual condition was actually the usual optimality condition for private goods extended to public goods. This condition stated that the cost of acquiring the collective good differs from one individual to the other and should correspond to the amount each individual is willing to pay for the good to guarantee a Pareto allocation of resources. Put differently, for each individual, the relative price of a public good in terms of a private good should equal the marginal rate of substitution between those two goods. To guarantee an optimal allocation of resources with private and public goods, Buchanan (1955a, 4; emphasis in original) wrote, "each individual must equate the marginal rate of substitution in consumption between any collective good and one private good with the marginal rate of substitution between these two goods in production *to him*." Then, each individual would be certain to be on "his own utility frontier" (Buchanan to Samuelson, March 4, 1955, BP).

The individual condition—Buchanan admitted—would be "redundant" (Buchanan to Samuelson, February 25, 1955, BP) or "not needed in the Samuelson formulation" (Buchanan 1955a, 7). Thus, Samuelson's condition "provide[d] a pure theory of public *finance* in the Samuelson welfare economics" (2; emphasis in original)—that is, in a frame in which exists a social welfare function. Given such a function, no additional individual condition was needed to say how the costs of the provision of the public good should be distributed among individuals. Individual shares were given by, and incorporated in, the social welfare function. However, Buchanan's condition became necessary for those who like him were "not willing move beyond the 'narrow' or Paretian version of modern welfare economics" (Buchanan 1955a, 1). Buchanan's condition was an "alternative" to Samuelson's social welfare function. "This is all that it ever intended to be" (Buchanan to Samuelson, March 4, 1955, BP). It was the condition for an alternative welfare economics in which there was no social welfare function. Adding up the individual willingness to pay was necessary and sufficient to guarantee a Pareto optimal allocation of resources. All individuals would agree to pay what they were asked to pay, because this was what they wanted to pay—"[t]he amounts *actually* paid are made equal to the amounts *willingly* paid" (Buchanan to Samuelson, February 25, 1955, BP). In other words, unanimous consent would be reached. In Buchanan's welfare economics, unanimity replaced a Samuelson-style social welfare function.

Individualized Prices, Collective Action, and Clubs

Unanimity is not easy to reach, and Buchanan's condition was not easy to satisfy. The prices for public goods could be individualized only if individuals revealed their true preferences and willingness to pay for the good. This was also a further point of disagreement between Buchanan and most economists, starting with Samuelson and Richard Musgrave.

Musgrave believed that when faced with a collective action problem, individuals would not indicate any willingness to bear the implied costs. Instead, driven by their self-interest, they would free-ride. In the late 1930s, Musgrave (1939, 220; see also Musgrave 1959) spoke of "the absence of a general willingness to comply with the obligation to contribute." Fifteen years later, Samuelson (1954, 388–89) wrote in a similar spirit: "It is in the selfish interest of each person to give false signals, to pretend to have less interest in a given collective consumption activity than he really has." As a consequence, in the presence of public goods and externalities, the use of individualized prices and nongovernment or decentralized mechanisms was impossible. Such "a failure of market catallactics" included, as Samuelson (1954, 389) noted, "voting and signaling schemes—by which he meant "'Scandinavian consensus', Kant's 'categorical imperative,' other devices meaningful only under conditions of 'symmetry', etc."—that Samuelson found "utopian." From Samuelson's (1955, 356) view, government intervention was the only option to deal with externalities and public goods: "Myriad 'generalized external economy and diseconomy' situations . . . provide obvious needs for government activity" (for similar views at the time, see among others Brownlee and Heller 1956; Margolis 1957; Musgrave 1939, 1959; Wiseman 1957).

For his part, while Buchanan readily admitted that markets fail to allocate resources efficiently in the presence of public goods (1954a, b), he also believed that such failures did not indicate a need for government intervention. For one thing, there were many problems that individuals could not solve privately but that did not warrant collective action—"[t]he mere presence of public or collective needs [should not be] confused with the necessity for satisfying them . . . . The existence of 'undeniable' need does nothing toward proving that action must be taken to meet it" (Buchanan 1957b, 175; see also 1959; Buchanan and Tullock 1962; Buchanan and Stubblebine 1962; Buchanan and Kafoglis 1963). Buchanan, Wicksellian in his confidence, was convinced that at least under certain conditions individuals would cooperate to solve these market failures. In contrast to Samuelson, Musgrave, and many others, Buchanan claimed that individuals do not *always* free ride in the presence of public goods. Government intervention was required only when the voluntary individual efforts at collective action failed.

In Buchanan's view, the key issue for addressing issues of public goods, externalities, and other market failures was not to identify "those goods and services that exhibit 'publicness'" (1965b, 11). Instead, the first step was to start from the collective solutions individuals willingly devised. As a corollary, the role of the economist was to understand the conditions under which individuals would voluntarily act

collectively and cooperate to solve or prevent market failures without having to rely on government intervention. Economists must "concentrate their attention on a particular form of human activity and upon the various institutional arrangements that arise as a result of this form of activity" (Buchanan 1964a, 213–14). This was the point Buchanan made in his Nobel prize lecture (1987). This was the research program Buchanan outlined in "What Should Economists Do?" (1964a) and to which belonged "An Economic Theory of Clubs" (1965a; 1964b).[3] The goal was here to explain how individuals devise "consumption ownership-membership arrangements" (Buchanan 1965a, 1), "cost and consumption sharing arrangement[s]" (Buchanan 1965a, 2), "membership or sharing arrangements" (Buchanan 1965a, 13) or, still in other words, "an organization of persons designed solely for the objective of utilizing a single communal community" (1964b).

Clubs, Small Numbers, and Property Rights

Buchanan developed his "theory of co-operative membership" (1965a, 1) over many articles and books—including, among others, "An Economic Theory of Clubs" (1965a), "Simple Majority Voting, Game Theory and Resource Use" (1961a) and "Ethical Rules, Expected Values, and Large Numbers" (1965b). Indeed, one major aspects of Buchanan's analysis related to the number of persons involved in the problem. More precisely, clubs exist when and because some individuals are willing to cooperate and to act collectively. Such a willingness, in turn, depends on the number of persons affected by the market failure and on the nature of the property rights involved.

To Buchanan, free riding and cooperation depended on whether or not individuals were in large or in small groups, because it depended on the probability each individual "assigns to the various patterns of behavior for 'others' than himself" (1965b, 5; 1968b, 85). The point was crucial because it had been neglected by economists (see also Buchanan 1978, 364–65). Numbers were crucial to mark a difference with the economists with whom Buchanan disagreed and to explain when clubs would exist. Thus, in large groups, Buchanan argued individuals follow their self-interest in the narrow sense of the word independently from how others behave. Each individual treats others as parts of the environment, assuming that their action cannot influence others' and hence not adapting their behaviors to what others do in a way that would allow collective action to emerge (Buchanan 1965b, 1968a). They face what Buchanan called "the large-number dilemma" (1965b, 1968b). Although individuals are aware that they would be better off by contributing to the provision of the public good or internalizing the external effects of their action, they nonetheless "refuse, rationally, to contribute to this cost on an individualistic and voluntary basis" (1965b, 9). In short, they behave as assumed by

[3] The concept of club goods was not totally new to Buchanan. For predecessors, see Wicksell (1896, 114; 1958, 90), Benham (1934, 451), and Wiseman (1957).

the standard Samuelson/Musgrave public finance doctrine. Collective action fails. A certain form of state intervention was inevitable.

The situation was radically different in small groups. In those environments, Buchanan argued individuals behave strategically, adapting their behavior to what others do and what they anticipate about the behavior of others. In a small group, individuals might in some cases behave morally and follow a Kantian rule of action—to act in accordance with the rules you would like to see followed by everyone (Buchanan 1961a, 1965b, 1968b). If individuals in this setting follow an ethical rule of action, they contribute to the provision of public goods, internalize the effects they have on others—say, in trying not to make too much noise in public spaces—or bargain with others over these external effects or, in addition, do not cheat on their commitments. To put it differently, they act collectively. To Buchanan, there was no doubt that in small groups, the probability to follow this Kantian principle was higher than in large groups (see also 1978, 365).

The willingness to cooperate in smaller groups, as likely as it was, was not guaranteed without certain rules. Individuals could indeed behave opportunistically, "find[ing] it to [their] advantage to conceal [their] true preferences and to give false signals about those preferences to [their] opponents-partners" (Buchanan 1968b, 81). Such behaviors could be tolerated only up to a certain point (Buchanan 1968a, 357–58). The danger that some individuals could free ride was an obstacle to collective action. Individuals would indeed "be reluctant to enter voluntarily into cost-sharing arrangements . . . [i]f they think that exclusion will not be fully possible, that they can expect to secure benefits as free riders without really becoming full-fledged contributing members of the club" (Buchanan 1965a, pp. 13–14). Free riders should be excluded from the group.

Exclusion—that is, the exclusion of those who do not pay the membership fees—therefore has to be possible. In small groups, where relationships were personal, "the possibility of *excluding* genuine non-conformists will normally be present" (Buchanan 1968b, 87). But it was much more efficient if property rights were used to guarantee the exclusion of free riders, to prevent free-riding behaviors, and therefore to secure the benefits of inclusion in the club to its members (Buchanan 1965a, 13; see also Buchanan and Tullock 1962, 44). Property rights were necessary to allow the exclusion of potential free riders.

A club was therefore viewed as an institutional arrangement devised to include the individuals who were willing to cooperate—that is ready to adopt an ethical rule of behavior—and to exclude those who could be tempted to free ride. As Buchanan (1965a, 13) wrote, a "theory of club is . . . a theory of optimal exclusion, as well as one of inclusion." As a mechanism, clubs could be used to deal with many instances of market failures. Local goods—a swimming pool, which was the example Buchanan took in his 1965 article or a highway network as in the preliminary 1964 version—come immediately to mind. Thus, clubs seem to be rather small groups. However, there was no conceptual reason why clubs could not exist to deal with less impure and less local public goods, and even with pure public goods in the standard Samuelsonian sense of the word—such as defense. Buchanan cited lighthouses or a vaccine

as examples of goods that could be dealt with through clubs, provided that property rights were defined to prohibit free riders from benefiting of the good. Potentially, to Buchanan (1965a, 13), there was no limit to physical excludability: "Physical exclusion is possible, given sufficient flexibility in property law, in almost all imaginable cases, including those in which the interdependence lies in the act of consuming itself."

Clubs, Individualized Prices, and Pareto Optimality

Once ethical and legal rules guaranteed that individuals would cooperate and pay the price to support the club, the next question obviously bore on how the costs of the public good would be shared among members. For Buchanan, what would be a club pricing mechanism? He answered the question in "An Economic Theory of Clubs" (1965b) and also in incomplete drafts available in the Buchanan Papers, which we treat here as a single manuscript (1964b).

Initially, Buchanan envisaged a system in which consumers were charged a two-part price. First, a charge per unit of consumption set at the level of the marginal cost. Second, a flat or fixed fee to cover the costs of the provision of the good or the difference between the marginal and average cost, given that public goods and club goods are frequently decreasing cost industries. Inspired, again, by Wicksell and his defense of the use of marginal cost pricing for decreasing cost public enterprises, Buchanan suggested to price highway services by using such a two-part tariff (1952, 1955b).[4] It would consist of a gasoline tax—the "rough equivalent to a mileage toll" (Buchanan 1952, 102)—and motor vehicle license fees—to include elements such as vehicle weight, the type of road used, and the time at which vehicle travels. Buchanan added, without giving any details, that "tax financing [should] be considered . . . to cover the total costs of construction and maintenance" (Buchanan 1952, 100).[5]

Defined in this way, the prices for highway services would vary from one individual to the other, which made sense for Buchanan since the benefits received by each individual were different. Another reason was that the costs each motorist generated, and that had to be covered by the price, were different. There were the costs of highway depreciation that depended on the type of vehicle used and, most importantly, the costs imposed on others. Indeed, Buchanan pointed out

[4] Wicksell argued that public enterprises should "charg[e] fees equal to the marginal costs of providing the service and making up the deficit by tax revenues" (Buchanan 1951, 174). The tax was raised to finance the deficit caused by the difference between the average and marginal cost that characterized such industries. They should be paid by the users of the public good, that is, by "the individuals who benefit from the proposed enterprise" (176; see also Buchanan 1948, 1949, 1952, 108). Other prominent economists of the time like Lerner (1944) and Hotelling (1938) argued that such taxes should be levied on all taxpayers—including nonusers—in a non-discriminatory way.

[5] In his early writing, Buchanan (1952, 102) ruled out highway tolls. The system would have to be "comprehensive and highly differentiated," and it would be "completely unworkable from an administrative point of view, and would be uneconomic besides."

using highways generated "spillover" effects. Thus, the user charge had to be set at the level of the marginal *social* cost. Users would also have to pay the "spillover" costs" resulting from "[t]he progressive deterioration in quality of highway service as congestion increases" and that were "represented in poorer service provided [sic] all users" (1952, 100). Thus, for instance, "The motorist who drives a new and efficient vehicle does "benefit" more from highways than does the motorist who drives the "Model A." But the latter should pay a higher price because he adds more to social cost" (Buchanan 1952, 106).

Price discrimination would ensure that only those willing to pay would actually demand road services. Others would exclude themselves from the use of roads and highways. User prices would efficiently restrict demand, which was necessary to solve the major problem of that congestion on highways. To Buchanan, congestion evidenced too high a demand for highway services. The key was to ration demand: "The answer to the whole highway problem lies in 'pricing' the highway correctly. The existence of congestion on our streets and highways is solely due to the fact that we do not charge high enough 'prices' for their use" (1955b, 14–15).

But eventually, Buchanan changed his mind and rejected the two-part tariff—which is quite ironical if one remembers that such tariff is commonly linked to a club pricing mechanism (Sandler and Tschirhart 1980, 1504). Instead, Buchanan's preferred club pricing should not include a charge per unit, a variable part. Buchanan even suggested that it was more useful to stop talking of prices and to refer to "shares" (1965a, 12): "Users pay a share in the common costs of providing the facility" (1964b). However, those shares were not the same for all consumers. Buchanan (1965a, 4) noted that, "[f]or simplicity, we may assume equal sharing" but immediately added that "this is not necessary for the analysis." Buchanan stuck to individualized prices. In clubs, price discrimination is possible (see also Wiseman 1957, 64). The price, or shares, paid in the club should differ from one individual to the other. The difference should reflect, first, the spillover costs that using a public good generates and, second, the willingness to pay for the good. These two principles remained important.

Buchanan also changed his mind about how to take congestion into account in clubs. Club members—motorists, for instance—should no longer be asked to pay for the spillover costs and club goods should no longer be priced on the marginal social cost: [t]he club approach . . . involves no attempt to impose a charge on users that reflects spillover congestion costs. (Buchanan 1964b).[6] The reason seemed to have been that the externalities generated by highway users were nonseparable and, as Otto Davis and Andrew Whinston had demonstrated, marginal cost pricing could be used only when technological externalities are separable (Davis and Whinston

[6]He had "come to the view that all the stuff on trying to 'price' highways by measuring marginal costs of congestion, a position that [he] firmly supported in past, is conceptually wrong because it is impossible." (Buchanan to Tolley, October 7, 1964, BP)

1962, 247).[7] If the spillover costs could thus no longer be included from the cost side, they should be taken into account from the benefit side.[8] They were no longer paid by those who create the spillover effects but by those who suffer from them—if they are willing to do so—as foregone benefits. Buchanan (1964b) now claimed: "The costs of congestion enter the analysis through their effects on the estimated benefits to be received by final consumers." Each club member is characterized by a certain willingness to pay for additional members—a "rate (which may be negative) at which the individual is willing to give up (accept) money in exchange for additional members in the sharing group" (Buchanan 1965a, 4). Each additional member generates costs once in the club. Thus, "when the marginal benefits that he secures from having an additional member . . . are just equal to the marginal costs that he incurs from adding a member . . . an individual attains full equilibrium in club size" (1965a, 5).

Pareto optimality requires that each share was based on the individual's willingness to pay for having additional members in the club as well as the willingness to pay for the good produced by the club. That was the second element that club pricing should include. Buchanan (1965a, 5) thus added another principle stating that, for each individual, the "marginal rate of substitution between goods X_j [the club good] and X_r [the numeraire good], in consumption, must be equal to the marginal rate of substitution between these same two goods in 'production' or exchange." This was exactly the condition he had used in his 1955 comment to Samuelson. It reflected his conviction that each individual should pay the price that the individual is willing to pay. Clubs were thus meant to implement this so-important condition. This was also clearly a reason for which club shares would be individualized and different from one individual to the other. Again, Buchanan was implying that his condition and only his could guarantee a Pareto-optimal allocation of resources.

Thus, Buchanan's clubs were a means to implement the prices individuals were ready to pay. The two dimensions—foregone benefits and effective benefits—guarantee that outcome. Complementarily, what Buchanan meant was that, without clubs for public goods, individuals would pay a price that does not satisfy their preferences. Clubs were a means to avoid coercion.

Conclusion

With "An Economic Theory of Clubs," Buchanan (1965a) was trying to do much more than just fill in the conceptual gap between the extremes of pure public goods and pure private goods. He was not even trying to define a category of goods.

[7] An externality is separable if the consumption or output of *i* does not affect the marginal utility or cost of *j*. Otherwise, it is non-separable.

[8] "[T]he use of price to restrict usage to some 'optimal' level of traffic remains relevant [but], we should, I now think, come at price differently, and not via the cost side at all" (Buchanan to Tolley, October 7, 1964).

He sought to develop a different form of welfare economics, in which there is no social welfare function and in which utility functions cannot be "read" by external observers, but where individual preferences can instead only be revealed by actions taken in response to prices. Buchanan adopted this perspective to analyze the pricing of public goods and to discuss clubs. Buchanan explicitly presents his clubs as a means to *replace* Samuelson's condition for optimal spending on public goods (1965a, 6). Indeed, Buchanan's clubs are foreign and incompatible with the role that club goods play in traditional Samuelson-style public economics.

■ *I am grateful to Peter J. Boettke, David Coker, Andrew Farrant, Jean-Baptiste Fleury, and the editors of the Journal for very thoughtful comments.*

References

Benham, Frederic C. 1934. "Notes on the Pure Theory of Public Finance." *Economica* 1 (4): 436–58.

Boettke, Peter J., and Alain Marciano. 2017. "The Distance between Buchanan's 'An Economic Theory of Clubs' and Tiebout's 'A Pure Theory of Local Public Expenditures.' New Insights Based on an Unpublished Manuscript." *European Journal of the History of Economic Thought* 24 (2): 205–37.

Brownlee, Oscar H., and Walter W. Heller. 1956. "Highway Development and Financing." *American Economic Review* 46 (2): 232–50.

Buchanan, James M. 1948. *Fiscal Equity in a Federal State.* Chicago: University of Chicago.

Buchanan, James M. 1949. "The Pure Theory of Government Finance: A Suggested Approach." *Journal of Political Economy* 57 (6): 496–505.

Buchanan, James M. 1951. "Knut Wicksell on Marginal Cost Pricing." *Southern Economic Journal* 18 (2): 173–78.

Buchanan, James M. 1952. "The Pricing of Highway Services." *National Tax Journal* 5 (2): 97–106.

Buchanan, James M. 1954a. "Resource Allocation and the Highway System." Unpublished, BP.

Buchanan, James M. 1954b. "Consumption Interdependence and the Interpretation of Social Cost." Unpublished, BP.

Buchanan, James M. 1955a. "A Note of the Pure Theory of Public Expenditure." Unpublsihed, BP, reprinted in P. J. Boettke and A. Marciano, 2020, *The Soul of Classical Political Economy: James M. Buchanan from the Archives*, Mercatus Center.

Buchanan, James M. 1955b. *Traffic, Tolls and Taxes. The Economics of the Nation's Highway Problem.* Unpublished, BP.

Buchanan, James M. 1957a. "The Pure Theory of Local Expenditure: Comment." Unpublished, BP.

Buchanan, James M. 1957b. "Federal Expenditure and State Functions." In *Federal Expenditure Policy for Economic Growth and Stability*, 174–79. Washington, DC: United States Government Printing Office.

Buchanan, James M. 1959. "Positive Economics, Welfare Economics, and Political Economy." *Journal of Law and Economics* 2: 124–38.

Buchanan, James M. 1961a. "Simple Majority Voting, Game Theory and Resource Use." *Canadian Journal of Economics and Political Science* 27 (3): 337–48.

Buchanan, James M. 1961b. "Comments." In *Public Finances: Needs, Sources, and Utilization*, edited by James Buchanan, 122–29. Princeton: Princeton University Press.

Buchanan, James M. 1964a. "What Should Economists Do?" *Southern Economic Journal* 30 (3): 213–22.

Buchanan, James M. 1964b. "The "Club" Approach to Highways." mimeo, BP.

Buchanan, James M. 1965a. "An Economic Theory of Clubs." *Economica* 32 (125): 1–14.

Buchanan, James M. 1965b. "Ethical Rules, Expected Values, and Large Numbers." *Ethics* 76 (1): 1–13.

Buchanan, James M. 1968a. "A Behavioral Theory of Pollution." *Economic Inquiry* 6 (5): 347–58.

Buchanan, James M. 1968b. *The Demand and Supply of Public Goods.* Indianapolis: Liberty Fund, 1999.

Buchanan, James M. 1969. *Cost and Choice. An Inquiry in Economic Theory.* Chicago: Markham Publishing. Co.

Buchanan, James M. 1978. "Markets, States, and the Extent of Morals." *American Economic Review* 68 (2): 364–68.

Buchanan, James M. 1987. "The Constitution of Economic Policy." American Economic Review 77 (3): 243-50.

Buchanan, James M. 2007. *Economics from the Outside In: "Better Than Plowing" and Beyond*. Texas A&M University Press.

Buchanan, James M., and Charles J. Goetz. 1972. "Efficiency Limits of Fiscal Mobility: An Assessment of the Tiebout Model." *Journal of Public Economics* 1 (1): 25–43.

Buchanan, James M., and Milton Z. Kafoglis. 1963. "A Note on Public Goods Supply." *American Economic Review* 53 (3): 403–14.

Buchanan, James M., and William C. Stubblebine. 1962. "Externality." *Economica* 29 (116): 371–84.

Buchanan, James M., and Gordon Tullock. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Davis, Otto A., and Andrew Whinston. 1962. "Externalities, Welfare, and the Theory of Games." *Journal of Political Economy* 70 (3): 241–62.

Hotelling, Harold. 1938. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." *Econometrica* 6 (3): 242–69.

Johnson, Marianne. 2010. "Wicksell's Social Philosophy and His Unanimity Rule." *Review of Social Economy* 68 (2): 187–204.

Lerner, Abba P. 1944. *The Economics of Control: Principles of Welfare Economics*. New York: Macmillan.

Marciano, Alain. 2020. "How Wicksell became important for Buchanan: a historical account of a (relatively) slow epiphany." Journal of Public Finance and Public Choice, 35 (2): 181-203.

Margolis, Julius. 1957. "Welfare Criteria, Pricing, And Decentralization of a Public Service." *Quarterly Journal of Economics* 71 (3): 448–63.

Musgrave, Richard Abel. 1939. "The Voluntary Exchange Theory of Public Economy." *Quarterly Journal of Economics* 53 (2): 213–37.

Musgrave, Richard A. 1959. *The Theory of Public Finance*. New York: McGraw Hill Book Co.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36 (4): 387–89.

Samuelson, Paul A. 1955. "Diagrammatic Exposition of a Theory of Public Expenditure." *Review of Economics and Statistics* 37: 350–56.

Sandler, Todd. 2013. "Buchanan clubs." *Constitutional Political Economy* 24 (4): 265–84

Sandler, Todd, and John T. Tschirhart. 1980. "The Economic Theory of Clubs: An Evaluative Survey." *Journal of Economic Literature* 18 (4): 1481–1521.

Sandler, Todd, and John T. Tschirhart. 1997. "Club theory: Thirty Years Later." *Public Choice* 93: 335–55.

Tiebout, Charles. M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 64 (5): 416–24.

Wicksell, Knut. 1896. *Finanztheoretische Untersuchungen*. Jena: Gustav Fischer.

Wicksell, Knut. 1958. "A New Principle of Just Taxation." (1896 J. M. Buchanan, Trans.). In *Classics in the Theory of Public Finance,* edited by R. Musgrave and A. Peacock, 72–118. New York: St. Martin's Press.

Wiseman, Jack. 1957. "The Theory of Public Utility Price-An Empty Box." *Oxford Economic Papers* 9 (1): 56–74.

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 257–264

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by e-mail at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105.

Smorgasbord

The UK government has published a 600-page report of an independent commission led by Partha Dasgupta, *The Economics of Biodiversity: The Dasgupta Review* (February 2021, https://www.gov.uk/government/publications/final-report-the-economics-of-biodiversity-the-dasgupta-review). "Not so long ago, when the world was very different from what it is now, the economic questions that needed urgent response could be studied most productively by excluding Nature from economic models. At the end of the Second World War, absolute poverty was endemic in much of Africa, Asia, and Latin America; and Europe needed reconstruction. It was natural to focus on the accumulation of produced capital (roads, machines, buildings, factories, and ports) and what we today call human capital (health and education). To introduce Nature, or natural capital, into economic

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.257.

models would have been to add unnecessary luggage to the exercise. Nature entered macroeconomic models of growth and development in the 1970s, but in an inessential form. . . . We may have increasingly queried the absence of Nature from official conceptions of economic possibilities, but the worry has been left for Sundays. On week-days, our thinking has remained as usual. . . . [I]n recent decades eroding natural capital has been precisely the means the world economy has deployed for enjoying what is routinely celebrated as 'economic growth' . . . If, as is nearly certain, our global demand continues to increase for several decades, the biosphere is likely to be damaged sufficiently to make future economic prospects a lot dimmer than we like to imagine today. What intellectuals have interpreted as economic success over the past 70 years may thus have been a down payment for future failure. It would look as though we are living at the best of times and the worst of times."

A National Academies of Sciences report investigates *High and Rising Mortality Rates Among Working-Age Adults* (March 2021, https://www.nap.edu/catalog/25976/high-and-rising-mortality-rates-among-working-age-adults). "The committee identified three categories of causes of death that were the predominant drivers of trends in working-age mortality over the period: (1) drug poisoning and alcohol-induced causes, a category that also includes mortality due to mental and behavioral disorders, most of which are drug- or alcohol-related; (2) suicide; and (3) cardiometabolic diseases. The first two of these categories comprise causes of death for which mortality increased, while the third encompasses some conditions (e.g., hypertensive disease) for which mortality increased and others (e.g., ischemic heart disease) for which the pace of declining mortality slowed. . . . [I]ncreasing mortality among U.S. working-age adults is not new. The committee's analyses confirmed that a long-term trend of stagnation and reversal of declining mortality rates that initially was limited to younger White women and men (aged 25–44) living outside of large central metropolitan areas (seen in women in the 1990s and men in the 2000s), subsequently spread to encompass most racial/ethnic groups and most geographic areas of the country. As a result, by the most recent period of the committee's analysis (2012–2017), mortality rates were either flat or increasing among most working-age populations. Although this increase began among Whites, Blacks consistently experienced much higher mortality."

Another National Academy of Sciences report considers *The Future of Electric Power in the United States* (2021, https://www.nap.edu/catalog/25968/the-future-of-electric-power-in-the-united-states). "[T]he committee identified a number of driving forces—social, technical, economic—that are likely to alter the landscape of the U.S. power system. These include the following: 1. Possible large growth in future demand for electricity. 2. Efforts to decarbonize the U.S. economy, and eliminate the emission of conventional pollutants, both by transitioning power generation to low or zero-emission sources and by making much greater use of decarbonized electricity as a substitute for fossil fuels in transportation, buildings and industry. 3. Developments at the edge of the grid such as distributed generation, storage, microgrids, energy management resources, and energy efficiency measures. 4. Grid stability challenges arising as a result of high penetrations of

nondispatchable sources of generation such as wind and solar. 5. A desire to reduce social inequities. 6. Concerns about the impacts of the energy transition on employment. 7. A changing international environment including powerful market forces arising from globalization, shifts in the locus of electricity-relevant innovation, and growing concerns about state-sponsored competition and disruption."

W. Brian Arthur offers a personal overview of the "Foundations of complexity economics" (*Nature Reviews Physics* 3: 136–45, 2021, https://www.nature.com/articles/s42254-020-00273-3). "Complexity economics sees the economy—or the parts of it that interest us—as not necessarily in equilibrium, its decision makers (or agents) as not superrational, the problems they face as not necessarily well-defined and the economy not as a perfectly humming machine but as an ever-changing ecology of beliefs, organizing principles and behaviours. . . . A new theoretical framework in a science does not really prove itself unless it explains phenomena that the accepted framework cannot. Can complexity economics make this claim? I believe it can. Consider the Santa Fe artificial stock market model."

The 2021 *World Development Report*, an annual flagship reports of the World Bank, is focused on the theme of "Data for Better Lives" (March 2021, https://www.worldbank.org/en/publication/wdr2021). From the weblink: "Today's unprecedented growth of data and their ubiquity in our lives are signs that the data revolution is transforming the world. And yet much of the value of data remains untapped. Data collected for one purpose have the potential to generate economic and social value in applications far beyond those originally anticipated. But many barriers stand in the way, ranging from misaligned incentives and incompatible data systems to a fundamental lack of trust. *World Development Report 2021: Data for Better Lives* explores the tremendous potential of the changing data landscape to improve the lives of poor people, while also acknowledging its potential to open back doors that can harm individuals, businesses, and societies."

Luís Brandão-Marques, Marco Casiraghi, Gaston Gelos, Günes Kamber, and Roland Meeks discuss the experience of "Negative Interest Rates: Taking Stock of the Experience So Far" (IMF Monetary and Capital Markets Department, 21-03, March 2021, https://www.imf.org/en/Publications/Departmental-Papers-Policy-Papers/Issues/2021/03/01/Negative-Interest-Rates-Taking-Stock-of-the-Experience-So-Far-50115). "Overall, most of the theoretical negative side effects associated with NIRP [negative interest rate policies] have failed to materialize or have turned out to be less relevant than expected. Economists and policymakers have identified a number of potential drawbacks of NIRP, but none of them have emerged with such an intensity as to tilt the cost-benefit analysis in favor of removing this instrument from the central bank toolbox. . . . [O]verall, bank profitability has not significantly suffered so far . . . and banks do not appear to have engaged in excessive risk-taking. Of course, these side effects may still arise if NIRP remains in place for a long time or policy rates go even more negative . . . The literature so far has largely overlooked the impact of negative interest rates on financial intermediaries other than banks."

Shawn Sprague dissects "The U.S. productivity slowdown: an economy-wide and industry-level analysis" (*Monthly Labor Review*, April 2021, https://www.bls.

gov/opub/mlr/2021/article/the-us-productivity-slowdown-the-economy-wide-and-industry-level-analysis.htm). "The figure—$10.9 trillion—represents the cumulative loss in output in the U.S. nonfarm business sector due to the labor productivity slowdown since 2005, also corresponding to a loss of $95,000 in output per worker. . . . [N]ot only has the productivity slowdown been one of the most consequential economic phenomena of the last two decades, but it also represents the most profound economic mystery during this time, and though many economists have grappled with the issue for over a decade and even created some innovative research approaches to address the question, we still cannot fully explain what brought on this situation. . . . This article presents two approaches to address these questions . . . First, the economy-wide slowdown in labor productivity growth is analyzed by breaking out the series into its three component series: multifactor productivity (MFP) growth, the contribution of capital intensity, and the contribution of labor composition. Second, industry-level productivity data are used to identify the industries that made notable contributions to the economy-wide labor productivity slowdown."

Symposia and Books

The April 2021 issue of the *Southern Economic Journal* begins with the Presidential Address of W. Kip Viscusi to the Southern Economic Association on "Economic lessons for COVID-19 pandemic policies" (pp. 1064–89, https://onlinelibrary.wiley.com/toc/23258012/2021/87/4). It then includes an 11-paper symposium on "The Political Economy of the COVID-19 Pandemic." Viscusi writes: "Given the tremendous benefits that could be derived by having more adequate medical resources, it is preferable from a benefit-cost standpoint to make provisions before health crises arise so that severe rationing is not required for the next pandemic. In anticipation of future pandemics, it is feasible to acquire high-quality ventilators at a cost from $25,000 to $50,000. Adding in the cost of medical support personnel would raise the annual cost to about $100,000. A reserve supply of ventilators could be a component of an anticipatory pandemic policy. Preparing for future pandemics remains a cost-effective strategy even for annual probabilities of a pandemic on the order of 1/100. However, survey evidence by Pike et al. (2020) suggests that support for protective efforts of this type is unlikely to emerge, as there is a lack of public concern with long-term pandemic risks. As a result, there is likely to be a continued shortfall in preparations for prospective risks, leading to future repetitions of the difficult rationing decisions posed by COVID-19. . . . If human life is accorded an appropriate monetized value, the application of VSL and efficient principles for controlling risks will lead to greater levels of protection than will result if medical personnel follow the guidance provided by many prominent medical ethicists."

In the lead paper in the symposium that follows, Peter Boettke and Benjamin Powell describe "The political economy of the COVID-19 pandemic"

(pp. 1090–1106). "[F]rom the perspective of promoting overall societal well-being, we believe that governments in the United States and around the world made significant errors in their policy response to the COVID-19 pandemic. . . . The activities of the young and healthy impose a negative health externality on the old and infirm. But it is equally true that if the activities of the young are restricted because of the presence of the old and infirm, this latter group has imposed a negative externality on the young and healthy. If transactions costs were low, the Coase theorem would dictate that it would not matter to which party the rights to activity or restriction were assigned, as bargaining would reach the efficient outcome. However, in the case of COVID-19, and large populations, it is quite clear that transactions costs of bargaining would be prohibitive. Thus, the standard law and economics approach would recommend assigning rights such that the least cost mitigator bears the burden of adjusting to the externality. In the case of COVID-19, it is clear that the low opportunity cost mitigators are the old and infirm. Thus, Coasean economics would recommend allowing the activities of the young and healthy to impose externalities on the old and infirm, not the other way around. Lockdowns and stay at home orders get the allocation of rights exactly backwards and result in large inefficiencies because costs are disproportionately borne by the high cost mitigators."

Monica de Bolle, Maurice Obstfeld, and Adam S. Posen have edited 12-chapter e-book titled *Economic Policy for a Pandemic Age: How the World Must Prepare* (Peterson Institute for International Economics, April 2021, https://www.piie.com/publications/piie-briefings/economic-policy-pandemic-age-how-world-must-prepare). As one example, Martin Chorzempa and Tianlei Huang describe "Lessons from East Asia and Pacific on taming the pandemic": "Bloomberg News' COVID Resilience Rankings evaluate success in handling the pandemic while minimizing the impact on business and society. An astounding ten of the top 15 countries and territories are in East Asia and Pacific. Top performers vary enormously in size, wealth, and political institutions, from small, wealthy, democratic islands like Taiwan and New Zealand to large, middle-income countries under one-party rule like mainland China and Vietnam. Core to their exemplary performance was the use of targeted and less costly mitigation measures that do not require an economic freeze. . . . The experience in East Asia and Pacific varies among countries with diverse cultures, geographies, and political systems, but one thing is clear: rigorous masking requirements, testing, contact tracing, selective quarantines, border closings, and clear public health communication all helped to avoid the overwhelming economic dislocations that occurred in the West. . . . One of the most crucial advantages in the early days of a pandemic is testing capacity, which helps identify both individuals to quarantine and where to focus further testing."

Donald J. Boudreaux and Randall G. Holcombe have written *The Essential James Buchanan*. (Fraser Institute, May 2021, https://www.fraserinstitute.org/studies/essential-james-buchanan). "Buchanan called such aggregative thinking the 'organismic' notion of collectives—that is, the collective as organism. From the very start, nearly all of Buchanan's lifetime work was devoted to replacing the organismic approach with the individualistic one—a way of doing economics and political

science that insists that choices are made, and costs and benefits are experienced, only by individuals. . . . The point is that exchange possibilities are not confined to the simple bilateral exchanges on which economists traditionally focus nearly all of their attention. When this truth is recognized, many familiar features of the real world are seen in a more revealing light. Clubs, homeowners' associations, business firms, churches, philanthropic organizations—these and other voluntary associations are arrangements in which individuals choose to interact and exchange with each other in ways more complex than simple, one-off, arm's length, bilateral exchanges. These 'complex' exchange relationships are an important reality for economists to study. But they are more than mere subject matter for research. They are also evidence that human beings who are free to creatively devise and experiment with alternative organizational and contractual arrangements have great capacity to do so. Where the conventional economist sees 'market failure,' humans on the spot often see opportunities for mutually advantageous exchange."

Interviews

Douglas Clement provides an "Esther Duflo interview: Deciding how to share" (*For All*: Federal Reserve Bank of Minneapolis, Spring 2021, https://www.minneapolisfed.org/article/2021/esther-duflo-interview-deciding-how-to-share). On the a tradeoff between growth and inequality: "I think the whole notion of a trade-off is likely a fallacy, for various reasons. First of all, there is no clear link either on theoretical grounds or empirically between higher inequality and more growth. There is no reason why inequality is necessary for growth. And there is no law of economics that says that growth increases inequality either. So I think there is no causality necessarily going in either direction; therefore, there is not necessarily a trade-off. Just as a matter of accounting, growth is equality-enhancing if most of the benefits of growth are going toward the poor. And growth is inequality-enhancing if most of the advantages are going toward the rich. Both are possible. I don't think there is a systematic pattern either way. . . . In fact, we don't seem to have much of a handle on what causes growth anyway, although we might have interesting theoretical narratives on growth. If there is a consensus among macroeconomists, it's on what should be avoided at all costs, like hyperinflation. But there is not a set of recipes that guarantees growth, and it's not that these recipes therefore lead to a trade-off."

Michael Chui and Anna Bernasek of the McKinsey Global Institute interview Christopher Pissarides "about how he developed the matching theory of unemployment, how COVID-19 affected his research, and what might be in store for labor markets after the pandemic" (May 12, 2021, "Forward Thinking on unemployment with Sir Christopher Pissarides," https://www.mckinsey.com/featured-insights/future-of-work/forward-thinking-on-unemployment-with-sir-christopher-pissarides, audio and transcript). "[B]efore we did that work, people were thinking of unemployment as a kind of stock of workers, as a number of workers if you like, who could not get a job. They would start from the top end of the market and say, 'This

is how much output this economy needs, that's how much is demanded. Then how many people do you need to produce that output?' Then you would come up with a number. And then they would say, 'Well, how many workers want jobs?' If there are more workers that want jobs, you call the difference unemployment. . . . What we did was to start from below, saying the outcomes in the labor market are the result of workers looking for jobs, companies looking for workers. The two need to come together. . . . [T]he time that it takes to find that job depends on how many jobs are being offered in the labor market, what types of skills firms want, what incentives the worker has to accept the jobs, what's the structure of production, the profit that the firm expects to make, conditions overall in the market. All those things influence the duration of unemployment. Therefore you could study there—how long does the worker remain unemployed? What could influence that duration? What could make it shorter? What would make it longer if you did certain things? On that basis, you derive good policies towards unemployment, and they are still the policies that governments use, in fact widely, to work out how long people remain unemployed and what the implications of their unemployment are."

David A. Price carries out an "Interview" with Matthew Jackson, with the subheading "On human networks, the friendship paradox, and the information economics of protest movements" (*Econ Focus*: Federal Reserve Bank of Richmond, 2021, Q1, pp. 16–20, https://www.richmondfed.org/publications/research/econ_focus/2021/q1/interview). "[O]ne key network phenomenon is known among sociologists and economists as homophily. It's the fact that friendships are overwhelmingly composed of people who are similar to each other. This is a natural phenomenon, but it's one that tends to fragment our society. When you put this together with other facts about social networks—for instance, their importance in finding jobs—it means many people end up in the same professions as their friends and most people end up in the communities they grew up in. From an economic perspective, this is very important, because it not only leads to inequality, where getting into certain professions means you almost have to be born into that part of society, it also means that then there's immobility, because this transfers from one generation to another. It also leads to missed opportunities, so people's talents aren't best matched to jobs." "This concerns another network phenomenon, which is known as the friendship paradox. It refers to the fact that a person's friends are more popular, on average, than that person. That's because the people in a network who have the most friends are seen by more people than the people with the fewest friends. On one level, this is obvious, but it's something that people tend to overlook. We often think of our friends as sort of a representative sample from the population, but we're oversampling the people who are really well connected and undersampling the people who are poorly connected. And the more popular people are not necessarily representative of the rest of the population. . . . There have been instances where universities have been more successful in combating alcohol abuse by simply educating the students on what the actual consumption rates are at the university rather than trying to get them to realize the dangers of alcohol abuse. It's powerful to tell them, 'Look, this is what

normal behavior is, and your perceptions are actually distorted. You perceive more of a behavior than is actually going on.'"

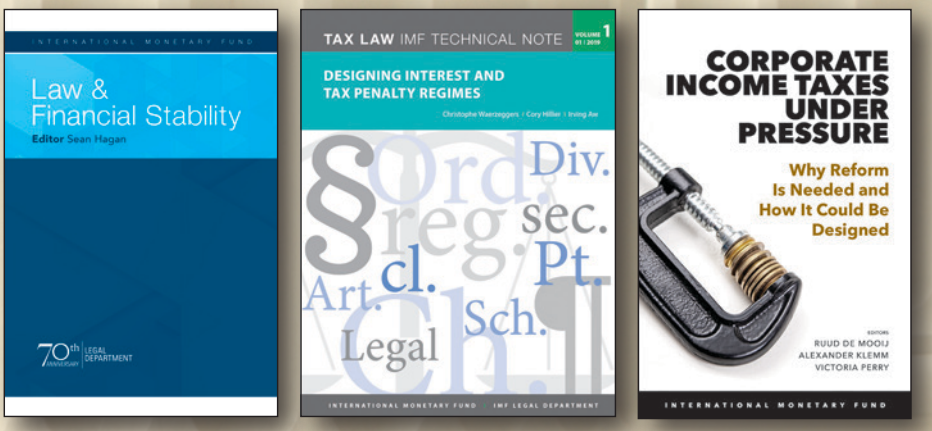
Discussion Starters

Michael Giberson considers "Texas Power Failures: What Happened in February 2021 and What Can be Done" (Reason Foundation, April 2021, https://reason.org/wp-content/uploads/texas-power-failures-what-happened-what-can-be-done.pdf). "The temperature in Dallas dipped to –2° F, the coldest it had been in Dallas for 70 years. Snow fell on the beaches on the Gulf Coast at Galveston, south of Houston. Temperatures in Austin remained below freezing for six days at a time of when temperatures usually average in the mid-50s. At Brownsville, near the most southern tip of Texas, February weather typically averages 65° F. High temperatures in Brownsville were in the mid-80s just days before the cold. . . . For the first time in history all 254 counties in Texas were under a winter storm warning at the same time. The cold was not unprecedented at any particular location, but it was extreme, widespread, and long lasting in February 2021. . . . Natural gas production and distribution froze up. Municipal water mains froze in cities across the South. Ranchers in the Panhandle lost cattle to the cold. Citrus growers in South Texas saw damage to trees that may last for years. Roads were closed due to ice and storms. Failures were not solely an electric power industry concern or a natural gas failure. The cold was simply worse than almost anyone in Texas was prepared for. . . . Clearly, it was not negligent on ERCOT's part—and maybe anyone's part—to fail to anticipate such anomalous temperatures."

Rachel Soloveichik discusses "Including Illegal Market Activity in the U.S. National Economic Accounts" (*Survey of Current Business*, February 2021, https://apps.bea.gov/scb/2021/02-february/pdf/0221-illegal-activity.pdf). "[E]xpenditure shares for all three broad categories of illegal drugs grew rapidly after 1965 and peaked around 1980. In total, this analysis calculates that illegal drugs accounted for more than 5 percent of total personal consumption expenditures in 1980. This high expenditure share is consistent with contemporaneous news articles and may explain why BEA chose to study the underground economy in the early 1980s . . . [A]lso . . . illegal alcohol during Prohibition accounted for almost as large a share of consumer spending as illegal drugs in 1980 and changed faster. Measured nominal growth in 1934, the first year after Prohibition ended, is badly overestimated when illegal alcohol is excluded from consumer spending."

IMF eLibrary

Legal Essential Reading



One-click access to IMF research on legal issues

- Books, working papers, and policy guidance on tax law, labor, banking rules, anti-money laundering, and digital currencies
- The latest research from the International Monetary Fund looking at the intersection of law and economics
- Information and resources on financial stability and international monetary law

IMF eLibrary also offers 22,000+ publications, archival material going back to 1946, and other curated Essential Reading Guides on key topics such as Risk Management and COVID-19.

All Completely FREE

LEGAL

eLibrary.IMF.org

jumpstart your research on global issues

INTERNATIONAL MONETARY FUND

AMERICAN ECONOMIC ASSOCIATION

BEST PRACTICES FOR ECONOMISTS:

BUILDING A MORE DIVERSE, INCLUSIVE, AND PRODUCTIVE PROFESSION

A more diverse profession fosters a more vibrant discipline.

See practical suggestions and supporting research from the *AEA Task Force on Best Practices* regarding actions all economists can and should take. With intention, we CAN make change.

- **CONDUCTING RESEARCH**
- **SERVING AS COLLEAGUES**
- **WORKING WITH STUDENTS**
- **LEADING DEPARTMENTS AND WORKPLACES**

More information at

www.aeaweb.org/resources/bestpractices

AEA INITIATIVES FOR DIVERSITY AND INCLUSION

The American Economic Association is committed to the continued improvement of the professional climate in economics. In cooperation with key committees, the Association has launched several new initiatives to support and promote diversity and inclusion in our profession.

1 AEA Award for Outstanding Achievement in Diversity and Inclusion

This annual award will recognize departments and organizations that demonstrate outstanding achievement in diversity and inclusion practices. Focus will be on those applicants that take productive steps to establish new programs and procedures to create an inclusive environment, and to increase the participation of underrepresented racial/ethnic minorities, women, and LGBTQ+ individuals.

2 Departmental Seed Grants for Innovation in Diversity and Inclusion

These grants, in amounts up to $5,000, will be awarded to economics departments to help establish new bridge programs or training programs for underrepresented minorities (URM). For example, a department might create a mentoring program for URM graduate or undergraduate students, create opportunities for URM students to do meaningful research assistant work, or start a program allowing URM students who need additional preparation for graduate school to take a lighter class load in the first year or to take core economics courses over two years.

3 The Andrew Brimmer Undergraduate Essay Prize

Thanks to the generosity of an anonymous donor, this paper prize has been established in honor of Andrew Brimmer, the first African American to serve on the Board of Governors of the Federal Reserve. The annual award will be presented to an undergraduate student at a US-based institution of higher learning majoring in economics, political science, public policy, or related fields for the best essay on the "economic well-being of Black Americans." The winner will receive a check for $1,000 and a plaque from the president of the AEA.

4 URM Travel Grants

This award is open to junior economics faculty members from traditionally underrepresented groups in the economics profession. The grants will advance career and professional development by defraying the costs of travel, lodging, and conference registration to attend the annual ASSA Meeting.

5 Small Group Breakfast Meeting for URM

Each year at the ASSA Meeting there will be a breakfast held with scholars from underrepresented minorities and prominent economists in attendance. The goal is to allow URM scholars access to AEA journal editors, executive board members, thought leaders in specific areas of economics, or other economists for the purpose of addressing issues of access to journals, conferences, and networks that are often out of reach for URM scholars.

These initiatives are another important step in helping make our field accessible and welcoming to anyone with the interest and ability to make a career in it. Please help us share this information throughout the profession so we can all work together and continue to improve.

For more details and information regarding how to apply for any of these initiatives, please visit

https://www.aeaweb.org/go/diversity-initiatives

Did you miss a session at the ASSA Annual Meeting?

View videos of over 160 AEA sessions from the 2021 ASSA Annual Meeting, compliments of the American Economic Association!

Research on a wide variety of economics topics presented by speakers from all areas of the discipline.

Engaging topics cover post-pandemic outlooks, gender differences, diversity in economics, and many other current areas of research.

Special presentations include the AEA Distinguished Lecture, the AEA/AFA Joint Lecture, the AEA Awards Ceremony, and the Nobel Lecture honoring the 2019 laureates.

AMERICAN ECONOMIC ASSOCIATION

View sessions at

www.aeaweb.org/conference/2021/aea-session-recordings

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives.* Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at www.aeaweb.org/journals/jep.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP.* However, given

the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion.
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant managing editor, Alexandra Szczupak at a.szczupak@aeapubs.org. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

JEP is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.

2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.

3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $24.00, $34.00, or $44.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Summer 2021, Volume 35, Number 3