Does working from home work? A natural experiment from lockdowns

Shen, Lucas

National University of Singapore, Singapore, Singapore

November 2022

Does Working from Home Work? A Natural Experiment From Lockdowns

Lucas Shen[†]

First version: March, 2021

This version: November, 2022

Latest version here

Abstract

Using tracked changes from a large open-source software platform, this paper studies how working from home affected the output of individuals working in tech. The basis of the natural experiment comes from idiosyncratic and state-imposed workplace closures during the COVID-19 pandemic. I find a negative but almost-negligible change in individual-level output of 0.5 percent (standard error of 0.091 percent). Overall, and based on descriptive analyses of the time-stamped data, tracked changes in software development cadences approximate regular work activity and provide a useful avenue for future studies of work.

Keywords: Real-time data; Work from home; GitHub; Labor economics

JEL: C81, J01, J24, M54, O3

[†] I am grateful to the editors, Cezar Santos and Flavio Toxvaerd, and two anonymous referees for comments that lead to a much improved manuscript. I thank Katherine Caves, Huang Yuting, Giovanni Ko, Jing Zhi Lim, Katja Mann, Gaurav Sood, Tang Cheng Keat, and Walter Edgar Theseira who provided valuable feedback at various stages. Economies and territories are referred to as "countries" for convenience. Interpretations and errors are those of the author. For the online supplementary material to this research, click here (direct linking to figures, tables, and sections in the online appendix should work if the online appendix is named `ghwfh-online-appendix.pdf` in the same directory as this working paper draft). Asia Competitiveness Institute, Lee Kuan Yew School of Public Policy, National University of Singapore. Email: lucas@lucasshen.com

I Introduction

Working from home (WFH) has been a staple discussion on employees' productivity since the pandemic. Flexibility and autonomy are commonly cited reasons why WFH improves productivity (e.g., BBC 2020b; Khanna 2020). The company Fitjitsu believes that "increased autonomy offered to its workers will help to improve the performance of teams and increase productivity" (BBC 2020a). Less time commuting is also a commonly cited reason, although it conflates the intensive and extensive margin (Whiting 2020; Barrero et al. 2021). Yet another reason is office distractions (Banbury and Berry 1998).

For employees in information and tech-related industries, the argument narrows. Conventional wisdom suggests they are well-suited for remote work (e.g., Alipour et al. 2020; Gottlieb et al. 2020; Lerman and Greene 2020). However, recent press coverage reveals contrasting opinions (e.g., BBC 2020c). Some, like Fujitsu, suggest productivity gains (BBC 2020a,b; Whiting 2020). Others, the CEO of Netflix, for instance, claim that WFH "has no positive effects and makes debating ideas harder" (BBC 2020d).

This study assesses the extent to which WFH has led to measurable changes in the output of individuals working in tech. Previous studies on WFH focus on non-tech work contexts with routine tasks, finding an increase in WFH productivity (e.g., Bloom et al. 2015). However, the impact of WFH on non-routine tasks is less clear. Focusing on tech workers who deal with ad-hoc troubleshooting, problem-solving, and non-routine tasks can shed light on WFH output gains for the broader working population. In addition to being vital to the economy, these information and communications technology professionals, perfectly suited for remote work (Gottlieb et al. 2020), provide a useful bound on WFH output gains.

To quantify and identify the impact of WFH on the output of individuals, I construct a novel dataset by combining real-time data tracked changes in open-source software development with state-imposed lockdown data by geocoding individuals to regions. The timing of workplace closures worldwide, orthogonal to individual-level output, and the geographical dispersion of individuals in the sample provide a plausibly exogenous source of variation. This setting forms the basis of the natural experiment in this study. If WFH confers output gains, then the measured output of an individual should be higher during WFH periods.

To measure individual-level output, I start with a census of timestamped tracked changes in projects from GitHub, an open-source software platform. These tracked changes are designed with the intention of archiving snapshots of a project for version control. I demonstrate that these tracked changes have utility as a way to monitor the output of individuals insofar as the tracked changes constitute iterative and incremental contributions to a project.

Given the unconventional use of tracked changes in projects as a metric of output, I start by unpacking descriptive findings into four themes that provide some sense of the reliability of tracked changes as office work activity. First, the subset of tracked changes originating from users who report their companies reveals a set of well-known and traditional tech companies. Second, tracked changes have a cadence consistent with the five-day workweek, where changes are most frequent on the weekdays. In addition, tracked changes respond to salient federal and national holidays. Third, tracked changes have a cadence consistent with traditional office hours. Moreover, tracked changes are sensitive to lunch and dinner breaks at the end of a workday. Finally, I show that, on average, there is no discernible difference in the user-perceived quality of projects active before and after the pandemic. Although the tracked change capture only open-source projects, I argue later in Section II.C that open-source is of high quality and the default when developers look for solutions. Overall, these findings suggest that the real-time tracked changes from GitHub are a viable avenue for studies on output.

Using the variation in workplace closures across regions, I estimate the difference in tracked changes of individuals before and after WFH. Subject to the validity of the identifying assumption that the precise timing of state-imposed WFH, for regions that eventually imposed WFH, is uncorrelated with the output dynamics of individuals, I find a negative but modest reduction in tracked changes. Relative to the periods with no WFH, the largest estimated decline in tracked changes at the individual level during the required WFH periods is 0.5 percent.

Using tracked changes at the individual-project level to account for project-specific cadences in software delivery yields similar findings. The largest estimated decline in tracked changes is only 0.8 percent. One concern is that the small magnitude of the estimates is caused by attenuation bias coming from individual WFH periods not corresponding perfectly with their locations. Iteratively dropping individuals in the sample with progressively lower quality of region geocoding does not substantially affect the estimates.

A question of interest is whether the tracked changes come from employees at work. Later in Section II.C, I turn to developer surveys which suggest that people who write code purely as a hobby are rare. In Section IV, I address this concern analytically using two different datasets. The first dataset uses only users who list the organization they work at. The second dataset sieves out only tracked changes occurring during the workday office hours. In both cases, the estimates are fairly consistent with those from the main results. The observed change in output during lockdowns is negative but modest.

One shortcoming of the natural experiment employed in this study is that we cannot measure how many individuals in the sample switch to working from home once the lockdowns start. The estimates, therefore, fall under an intention-to-treat basis instead of the WFH treatment itself. In the discussion Section V.A, I argue that the main source of non-compliance, and therefore bias under the exclusion restriction violation, comes from individuals already working from home during normal times. I turn to survey numbers on pre-pandemic remote work to help bound the compliance rate. This exercise suggests that the estimated decline in output when WFH is small (less than a full one percent) even for compliance rates lower than what surveys suggest.

While the pandemic has gradually revealed broader concerns about the impact of WFH, such as with teamwork and communication (e.g., Bloom 2020; Ford et al. 2021; Forsgren 2020; Gibbs et al. 2021; McDermott and Hansen 2021), concerns regarding individual-level output persist in the impetus to return to offices (e.g., Barrero et al. 2021; Ozimek 2020; Yang et al. 2022; YouGov 2020). Moreover, concerns about teamwork and communication ultimately relate to output as a bottom line. Since the study uses the COVID-19 pandemic as an event study, the estimates cannot be interpreted causally as in randomized controlled trials (e.g., in Bloom et al. 2015; Emanuel and Harrington 2021). However, the setting provides evidence that the change in objective output metrics of individuals when WFH is negative but minimal.

This study focuses on a specific part of the labor market: programmers and people who write code. The findings, however, have broader implications. Software work is well-suited for remote work (Alipour et al. 2020; Gottlieb et al. 2020; Lerman and Greene 2020). Hence the switch to WFH constitutes less of an adjustment. If such individuals encounter a dramatic decline in output when shifting to WFH, what more for the broader working population who require a bigger adjustment to WFH. Fortunately, for managers and policymakers, this is not the case.

More generally, the study contributes to our understanding of changes in work patterns for occupations that do not deal with repetitive and transactional work in their day-to-day tasks. Software work involves problem-solving like other white-collar occupations. This is as opposed to work revolving around repetitive tasks, such as call center representatives as in Bloom et al. (2015).[1] Like many other professions, issues arise on an ad-hoc basis for people who write code in their work duties. Troubleshooting unexpected problems, liaising with peers and managers,

[1] Emanuel and Harrington (2021) obtain similar results as Bloom et al. (2015) with a similar experimental setting, although their study goes beyond WFH productivity impact.

and checking confusing documents for potential solutions are part of their day-to-day. The GitLab 2018 developer survey reports that unclear directions, changing project requirements, and unrealistic deadlines are top reasons for fruitless efforts. Another survey, the Stack Exchange 2019 survey, reports that having to attend meetings, having insufficient manpower, and a lack of support from management are top impediments to productivity. In this sense, occupations where the work involves writing code are not that different from the rest of the workforce, where WFH is a realistic option.

A key advantage of this study is that the output metrics are more objective. Many studies on the WFH impact rely on self-perceived changes in output, which is problematic (Uddin et al. 2022; Ralph et al. 2020). Self-perceived productivity, for example, may correlate more with manager appraisals (Baruch 1996) than with output. The natural experimental setting also avoids both the Hawthorne Effect and Goodhart's law, where individuals are aware their work is being tracked and manipulate the metrics accordingly to appear productive (Baltes and Diehl 2018; Chrystal and Mizen 2003; Goodhart 1984).

Considering that this study finds little change in output after lockdowns among people in tech, perhaps the problem is not whether workers can deliver on work tasks when WFH. Rather, the problem lies in monitoring output and incentivizing workers. Managers prefer staff to be in the office because working in office aids in monitoring efforts. This interpretation is consistent with the finding in this paper. For general work context where tasks are non-repetitive and involves problem solving, like in software work, WFH does not have the kind of varied and dramatic impact cited in the media (e.g. BBC 2020a,b,c; Whiting 2020; BBC 2020d).

In the discussion, I further place my findings in the context of key-related studies. One set of studies examines software developers' output which also relies on various records of tracked changes (e.g., Bao et al. 2020; Ford et al. 2021; Forsgren 2020; McDermott and Hansen 2021). Another set looks at the output of information workers since software exists to track what windows are active on work computers (e.g., Gibbs et al. 2021; Yang et al. 2022). Overall, they find a limited impact on output. While this study finds limited WFH impact on output, the study by Forsgren (2020) and McDermott and Hansen (2021) using the same data source find a tangible impact on the working hours of individuals after the pandemic, a finding shared by studies using other data sources (e.g., DeFilippis et al. 2022; Friedman 2020).

Other related studies include Choudhury et al. (2021) who estimate a substantial 4.4 percent increase in a *WFA* (work-from-anywhere) vs. WFH setting. More broadly, this paper contributes to the literature on remote working in the context of the pandemic, such as those on public goods contribution (Choudhury et al. 2020; Kummer et al. 2020; Ruprechter et al. 2021), changes in emails patterns (DeFilippis et al. 2020), uneven household costs (Stanton and Tiwari 2021), and employment and health impacts (Angelucci et al. 2020). Finally, a set of studies look into the share of jobs that can be done WFH (Alipour et al. 2020; Bartik et al. 2020; Bloom 2020; Brynjolfsson et al. 2020; Gottlieb et al. 2020).

The next Section II provides tracked changes from GitHub as a metric of output. Section III unpacks four set of descriptive findings. Section IV reports the results. Section V discusses the findings and limitations. Section VI concludes.

II Data and Background

II.A Approximating Productive Output

The ideal and canonical productivity metric has some units of output scaled by some units of input. The productivity metric in Bloom et al. (2015) and Emanuel and Harrington (2021), for instance, have this form. Their experimental context is with call-center operatives. The output unit is the number of calls completed, and the input unit is the number of hours clocked. However, this form of productivity

Table 1—*Work-From-Home (WFH) Coding from OxCGRT*

OxCGRT WFH indicator	Type	Description from Oxford's Blavatnik School of Government (Petherick et al. 2020)
0	Non-binding	No measures
1	Non-binding	Recommended closing (or recommended work from home)
2	Binding	Required closing (or work from home) for some sectors or categories of workers
3	Binding	Required closing (or work from home) for all-but-essential workplaces (e.g. grocery stores, doctors)

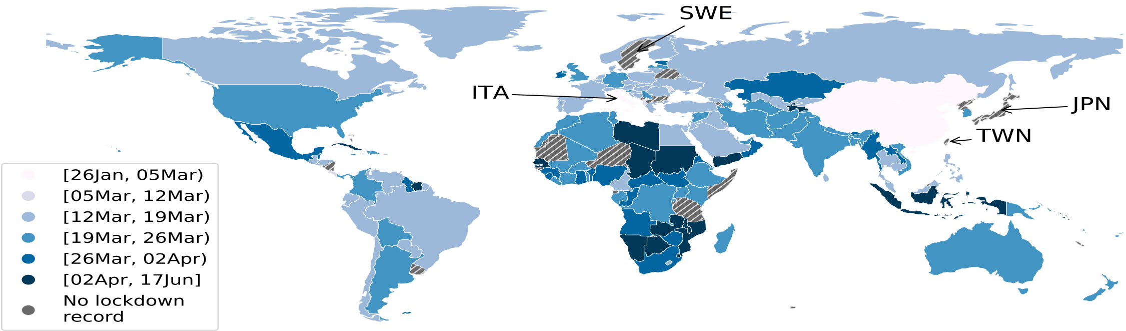


Figure 1: Staggered Timing in State-Imposed WFH

Notes—Map shows the timing of state-imposed WFH where workplaces are "required" to close. Underlying dates are for the date at which the OxCGRT WFH indicator first switches from {0, 1} to {2, 3}, that is, from no closing/recommended closing to a required closing for either some sectors or all-but-essential sectors (Table 1). Darker shades indicate later workplace closures. Countries without state-imposed WFH in the sample period (Jan–Jun 2020) are shaded in gray.

metric is rarely available outside of an experimental setting and in a work context that is less routine and transactional.

To gain empirical traction on the WFH impact outside of an experimental setting, and for work that is less routine, I use *tracked changes* on the GitHub platform as metrics of *output*.

GitHub is a platform where developers (and also some researchers) host, (Git) version control, and collaborate on projects. On the official website, GitHub states that it is "where the world builds software" and is the "largest and most advanced development platform in the world." The platform is free and open-source, with effectively zero barriers to entry. For a sense of scale, GitHub states that it has over 56 million users, 100 million repositories, and 3 million organizations. The users on GitHub are usually in the Information and Communications Technology (ICT) sector, and thus have work that is less routine than the work context in Bloom et al. (2015) for instance.[2] [3]

[2] Retrieved at time of writing from `https://github.com/about`. GitHub includes some of the most prominent organizations: Apple, Facebook, GitHub (dogfooding), Google, Microsoft, and Twitter, with repositories and users in the order of thousands. Google Maps, for instance, have several projects hosted on GitHub (`https://github.com/googlemaps`). See also Figure 4 and Figure 5

[3] See Papamichail et al. (2016) and Sanatinia and Noubir (2016) for a computing description of

II.B Tracking Changes in GitHub Projects

Using tracked changes of GitHub users in their repositories (projects) imply that we can track the activity of individuals. Here, I describe how we can approximate productive activity using the tracked changes from GitHub.[4]

In the Git workflow are two key milestones. First are commits, which are the first level of tracked changes in a project pipeline. These changes, for example, could be in a text file in the form of code or writing. When ready, users archive these changes, in a potentially modular fashion, to the local repository (which are then eventually pushed to the corresponding remote repository). To the extent that commits as tracked changes represent incremental improvements to code and, more generally, projects, I use commits as one metric of output.[5]

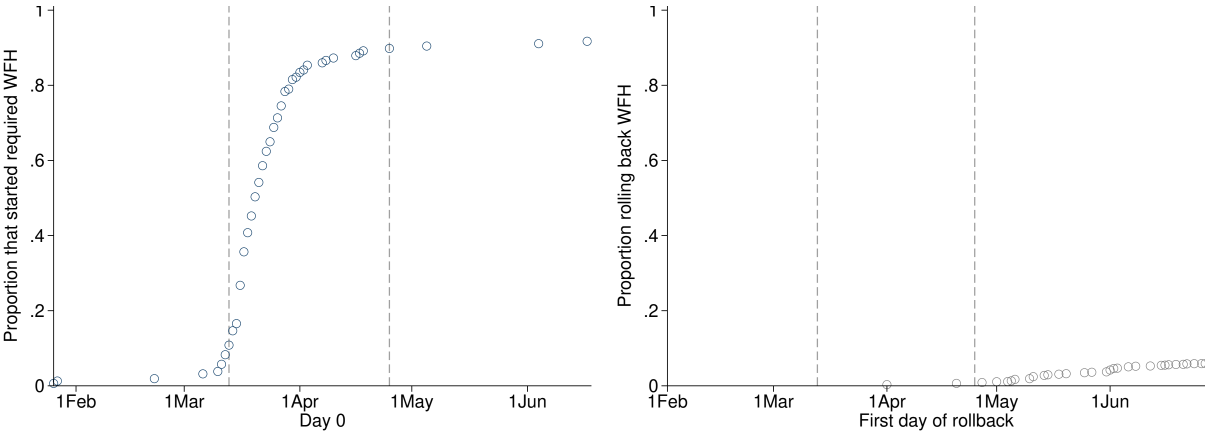
A second and usually larger milestone is a pull request. In the workflow, when an individual is happy with the (set of) changes—which could be a bug fix, issue resolution, or feature addition—they submit a pull request. Once the request is submitted, other members working on the same project can review and discuss and, upon approval, merge back to the main branch, which is always stable for production release.

Commits and pull requests are also metrics used by large tech companies such as Microsoft to monitor developer velocity (Spataro 2020). Commits are common metrics in modern software engineering (Baltes and Diehl 2018), and pull requests are a useful complementary metric since a single productivity metric is intrinsically problematic (Jaspan and Sadowski 2019). Figure A2 illustrates commits and pull requests as part of a branching workflow.

the GitHub platform.

[4] GitHub is not the only platform that allows researchers and managers to observe tracked changes in a context related to software development. Indeed, the studies by Bao et al. (2020) and Ford et al. (2021) also exploit tracked changes from partnering companies to study the WFH impact d the pandemic. I discuss these studies in Section V.B.

[5] In the related but separate relational database workflow, a "commit" is when a user commits computational resources to make a set of changes permanent and visible to other users in the pipeline. Changes from queries entered before committing are not reflected in the database.



(a) State-imposed WFH Timing

(b) Rollback Timing

Figure 2: Rollout of WFH

Notes—Panel (a) plots the cumulative proportion of countries implementing state-imposed WFH across dates in the sample period. Panel (b) plots the cumulative proportion of countries rolling back state-imposed WFH (Table A1 lists these countries that roll back state-imposed WFH in the sample period). The first and second vertical gray lines mark the dates at which approximately 10 and 90 percent of countries have imposed WFH.

It is worth noting two things about pull requests. First, they usually involve more serious projects, larger projects, and projects already in production, as the pull requests increase the barrier to making changes to the main pipeline. Pull requests also usually involves projects with collaborators.

In summary, I consider commits and pull requests as metrics of output and use them (separately) as the main outcome variables of interest for the rest of the study. I emphasize that GitHub does not track how long a user worked on a particular tracked change. Therefore, tracked changes are only metrics of productive output rather than the canonical form of productivity and should not be interpreted as such.

II.C Open-source Users and Projects

GitHub hosts open-source projects of all kinds and is open to all users. From the sample metadata, it is not possible to distinguish between commits and pull requests for work vs. those for hobby projects. However, according to the Stack Exchange (2022) survey, individuals who code only as a hobby and not work are rare, at less than 6 percent.

Using a survey (Zlotnick 2017) drawn from 5,500 GitHub users reveals that 70%

of respondents working on open-source repositories are employed. 85% of them contribute to open-source software in their day job (Section III.C later shows that open-source contributions follow regular day-job cadences). Most of these respondents (94%) are also end-users of open-source in their professional work, 81% use it frequently, and 65% contribute back to open-source as part of their work duties.

Overall, most people write code only when it is necessary for work. Section IV later shows that including only individuals who list their organizations, and only including tracked changes during hours consistent with a day-job, do not change the findings. This approach does not guarantee that the tracked changes exclude hobby projects but should better approximate tracked changes for the tasks at which individuals want to remain productive.

It is also worth noting that even open-source contributions that are not directly related to work duties can be productive for professional work. Virtually all professional developers have taught themselves a new language, framework, or tool. The Stack Exchange (2019) survey reports that 43 percent do so by contributing to open-source. This is 2.5 times more than those who go through full-time developer training bootcamps or industry certification programs. Other than online courses (presumably because of low entry costs), only on-the-job training has a higher proportion.

Finally, this study includes only open-source software and projects. However, the overwhelming consensus in the software development community is that the quality of open-source software is of the same, if not higher, quality than proprietary or closed-source software (Stack Exchange 2019). Open-source is the default when choosing software (GitLab 2018; Zlotnick 2017). The above factors suggest that contributions to open-source software are a viable avenue to study changes in individual-level output.

II.D Data Summary

To build the panel, I first query Google BigQuery's archive of GitHub timestamped commits in the period Jan 2020–Jun 2020, which includes author and repository names. Only public repositories are included; private repositories (opted out of public view) are not.

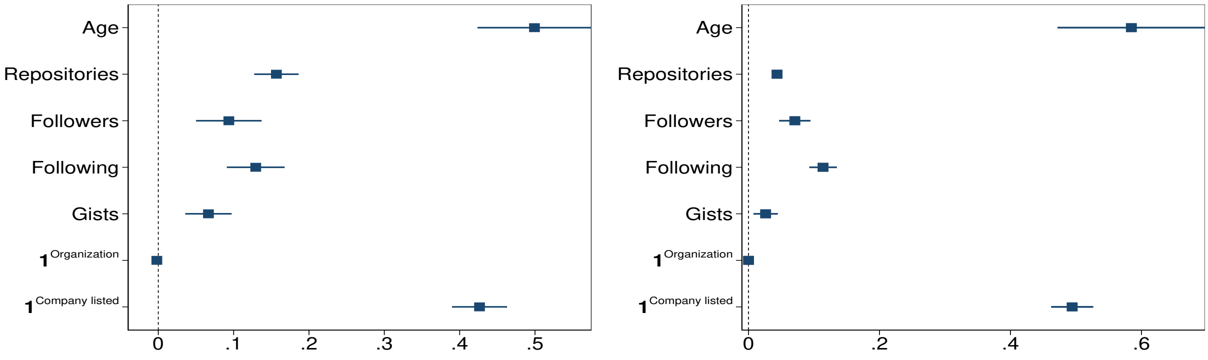
I use GitHub's Search API and User API to retrieve the location strings entered in the authors' user-profiles and then map users to countries by querying the `OpenStreetMap` API. Approximately half the users have a geolocation string in their user-profile, and most users with geolocation strings (98%) are successfully geocoded to a country. From this pipeline, I end up with approximately 300k users, 350k commit records, and 290k pull request records.[6]

To retrieve the WFH status for countries (and US states for the US) for any given date, I use the OxCGRT's repository of COVID-19 government responses trackers (Petherick et al. 2020). Table 1 lists the four types of OxCGRT WFH coding. Figure 1 shows the geographical variation in the start of state-imposed WFH, while Figure 2 shows the rollout of WFH across time.

For all the main analyses, I treat the recommended WFH from OxCGRT as non-binding while treating the two required WFH codings from OxCGRT (2 and 3) as binding and with a homogeneous effect. This assumes that the compliance of those working in ICT-related fields is the same for the two levels of state-imposed WFH. The assumption here is that software developers and researchers, and more broadly workers in the ICT sector, are one of the earliest workers who WFH during the pandemic. See also the studies by Alipour et al. (2020); Bartik et al. (2020); Bloom (2020); Brynjolfsson et al. (2020); Gottlieb et al. (2020) for the type/share of jobs that can WFH.[7]

[6] Appendix A.1 describes the data build in greater detail, while Appendix E of the Online Appendix provides randomly-sampled examples of both failed and successful geocoding, which I hand check. The formal results in Section IV include a robustness test for subsamples of the micro-level data depending on the quality of geocoding.

[7] In the event studies in the Online Appendix, for example, "Day 0" is defined as the date when



(a) Commits sample

(b) Pulls requests sample

Figure 3: GEOCODED VS OUT-OF-GEOCODED SAMPLE

Notes—Differences in means for geocoded records vs out-of-geocoded records (those not successfully geocoded, see Appendix E in the Online Appendix for examples). Units are in standard deviations (except the two indicator variables for scaling purposes only). Estimates derived from regressing the variables on the geocoded dummy and performing a t-test for the dummy. Number of individual observations are 44,894 and 120,614 for the commits and pulls request sample, respectively. Age refers to age of the individual's user account (creation date minus 1 Jan 2020). Repositories refer to the number of public repositories listed in the account. Followers and following are the number of accounts the individual follows and the number of accounts following the individual. Gists are the number of mini-blogs/code snippets. The last two dummies indicate whether the account type is an organization and whether the individual reports the company they work at. Tables A3 to A4 of the Online Appendix tabulates the above results. Robust standard errors clustered by countries (non-geocoded counts as a "country"). ***, **, and * denotes significance at the 1, 5, and 10 percent level, respectively.

III Descriptives

This section provides quantitative descriptions of commits and pull requests as tracked changes and unpacks four themes.

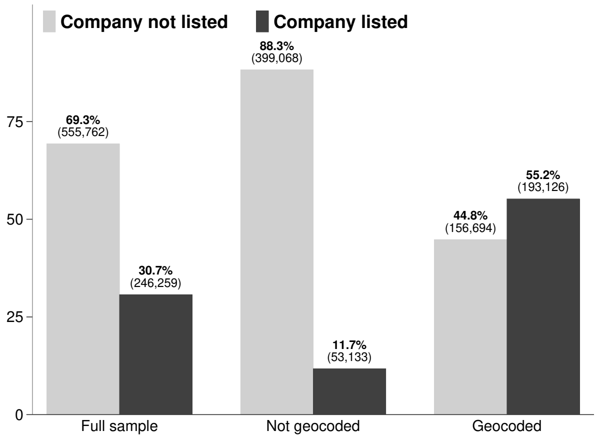
III.A Geocoded Individuals and Companies

The analyses below in Section IV include only GitHub users who have been successfully geocoded. So the natural question arises as to what separates users who have and have not been geocoded.

Figure 3 suggests that, by far, individuals who are successfully geocoded and therefore included in the analyses are the more prominent GitHub users. They have been on the platform for longer, have more projects, more followers, follow more people, and are more likely to report the company they work for (e.g., Microsoft). This stylized fact is true for both the commits sample and for the pull requests sample.

Figure 4 shows the proportion of commits that can be geocoded. In addition, the

the OxCGRT WFH indicator switches from a 0 or 1 to a 2 or 3.



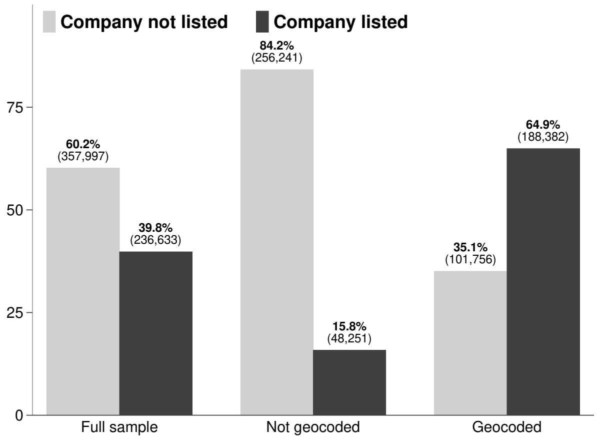
(a) Users listing company

	Company	Count		Company	Count
0	google	182	15	freelancer	13
1	microsoft	81	16	jetbrains	13
2	red hat	78	17	stanford university	13
3	alibaba	53	18	elastic	13
4	tencent	43	19	yandex	12
5	baidu	29	20	alipay	11
6	freelance	28	21	netease	11
7	bytedance	28	22	amazon	11
8	ibm	23	23	pingcap	11
9	facebook	20	24	shopify	10
10	automattic	18	25	esri	10
11	github	15	26	redhat	10
12	wso2	15	27	intel	10
13	thoughtworks	14	28	mercari	9
14	vmware	13	29	mit	9

(b) Most frequent companies

Figure 4: Distribution of Companies (Commits Sample)

Notes—Left panel shows the distribution of commits where the users have or have not self-report their companies on their user profile. Full sample is based on the total initial set of records. The geocoded sample is based on commits record where the user location is successfully geocoded. Right panel shows list of 30 most frequent appearing companies in the commits sample. Company names are derived using minimal preprocessing of the raw text.



(a) Users listing company

	Company	Count		Company	Count
0	google	949	15	linkedin	61
1	red hat	573	16	netflix	58
2	microsoft	500	17	automattic	57
3	ibm	157	18	thoughtworks	54
4	facebook	115	19	adobe	53
5	freelance	103	20	intel	50
6	vmware	98	21	amazon web services	50
7	mozilla	98	22	datadog	50
8	shopify	94	23	uber	49
9	github	91	24	alibaba	48
10	freelancer	84	25	aws	48
11	suse	74	26	elastic	47
12	tencent	63	27	salesforce	47
13	spotify	61	28	nvidia	45
14	hashicorp	61	29	yandex	45

(b) Most frequent companies

Figure 5: Distribution of Companies (Pulls Requests Sample)

Notes—Left panel shows the distribution of commits where the users have or have not self-report their companies on their user profile. Full sample is based on the total initial set of records. The geocoded sample is based on commits record where the user location is successfully geocoded. Right panel shows list of 30 most frequent appearing companies in the pull requests sample. Company names are derived using minimal preprocessing of the raw text.

figure shows the top thirty companies captured by the commits via the self-reported company of users. These companies are easily recognizable since they are some of the biggest tech companies.

Figure 5 shows the same picture for the pull requests sample. The list of most frequent companies is similar to those in the commits sample. What is different is that pull requests are more heavily represented by users who work for tech companies. Broadly, Figure 4 and Figure 5 suggest that the GitHub tracked changes can capture work activity from the well-known tech companies, and this is most

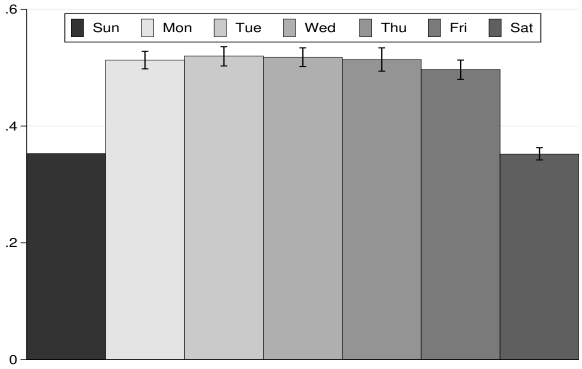


Figure 6: Day-of-Week Cadence *Notes—* Bar plots show the differences in log of (1+) commits by day-of-week (DoW) from regressing commits on the day-of-week dummies, plus user and repository fixed effects for the user-repository-DoW panel. The baseline day is Sunday—first bar—so that the standard errors for subsequent bars are for the estimates of the additional effects of Mon–Sat relative to Sunday. Robust standard errors are clustered at users and repositories.

pronounced with the pull requests since they constitute a key milestone in larger and collaborative workflows.

III.B Day-of-Week Patterns

Figure 6 shows systematically higher levels of commits on the weekdays, relative to Sunday. This implies most high-frequency output occurs on traditional workdays. This cadence manifests visually as an inverted U-shaped hump for commits from Sunday to Saturday.[8]

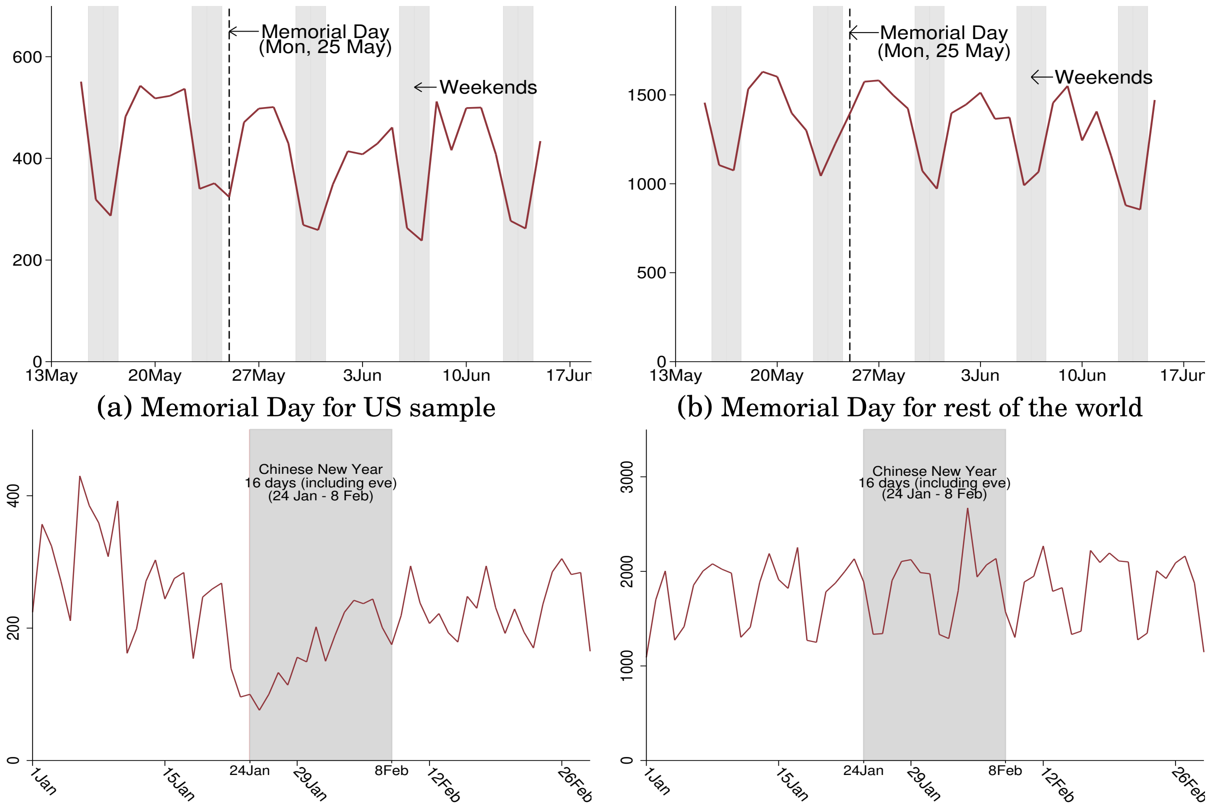
Figure 7 also shows the inverted U-shaped from the raw data. Panel (a) points out Memorial Day—a federal holiday in the US falling on a Monday in the US sample. The figure shows a distinct break in the day-of-week cadence since the relative inactivity over the weekend extends into Memorial Day on Monday before going back to normal on Tuesday. Notably, this break in the daily cadence is absent for the rest of the world in panel (b).

Panel (c) of Figure 7 shows how commits have a dramatic dip that directly coincides with the Chinese New Year holidays (between 2–7 holidays depending on region) for users geocoded to countries with high Chinese ethnic concentration. Panel (d) shows that the pronounced dip during the Chinese New Year is absent for

[8] To do this, I aggregate the commits log record up to the user-repository-DoW level, and then estimate

$$\ln(1+\text{commits})_{ijd} = \alpha + \sum_{d\in\{1,\ldots,6\}} \pi_d \text{DoW}_d + \text{individual}_i + \text{repository}_j + u_{ijd},$$

where Sunday ($d = 7$) is the reference day.



(a) Memorial Day for US sample

(b) Memorial Day for rest of the world

(c) Countries celebrating CNY

(d) CNY for rest of the world

Figure 7: Day-of-Week Cadence and Holidays

Notes—The first row shows the path plot (unsmoothed) of commits around May 2020 for US sample vs. the rest of the world. Black dashed line indicates Memorial Day on 25 May. Gray vertical bars indicate weekends (Sat–Sun). The second row shows the path plot (unsmoothed) of commits around Jan–Feb 2020, with the gray area indicating the Chinese New Year (CNY) period over a 15-days period plus the eve, 24 Jan–8 Feb. Panel (c) includes countries that celebrate the CNY: China, Indonesia, Korea, Malaysia, Singapore, and Vietnam. Majority of these are from China. Subfigure (d) includes users from the rest of the world.

the rest of the world.

Figures A14 to A15 in the Online Appendix make similar observations for the Martin Luther King Jr. Day for the US sample and the May Spring Bank Holiday in the UK. Broadly, the results from Figures 6 to 7 suggest that the metrics of output can indeed capture the expected day-of-week cadence as well as salient federal holidays.

III.C Time-of-Day Patterns

To push how far we can interpret commits and pull requests as metrics of output as office work activity, I start by converting the standard timestamps of commits to the local time. Figure 8 plots the density of commits across all 24 hours of a day

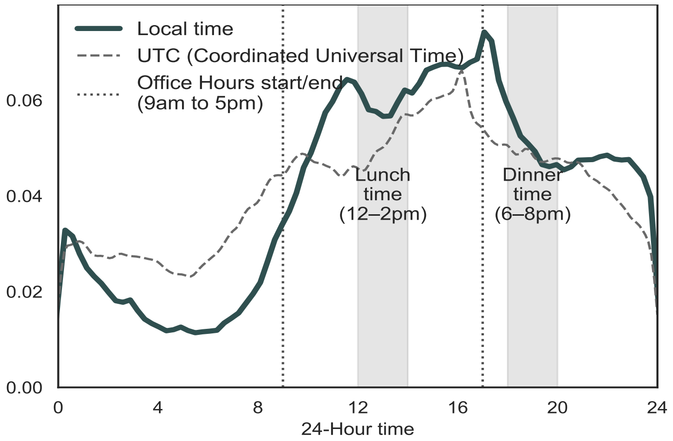


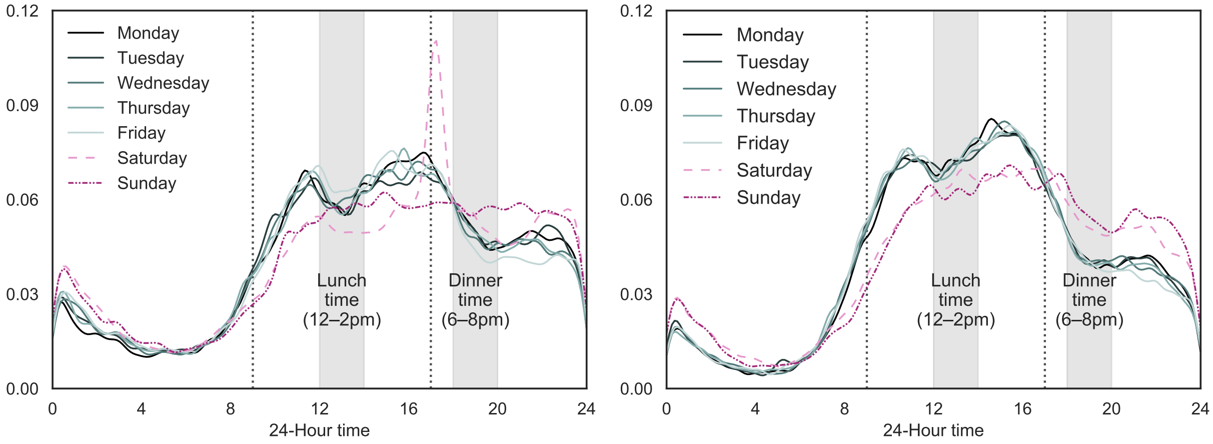
Figure 8: Time-of-Day Cadence
Notes—Kernel density estimate plot of the time-of-day-based cadence of the timestamped commits in 24-hour time. Only commits from users who are successfully geocoded are included. Minimal smoothing applied. Solid thick line is local time (UTC offset ± hh based on inferred local timezone). Thin gray dashed line is the timezone-agnostic timestamp from the commits records. The two vertical dotted lines are the start and end time of "office hour" (9am to 5pm). The two gray shaded areas indicate the two standard meal times (noon to 2pm for lunch time; 6pm to 8pm for dinner time).

based on the local time (inferred by matching self-reported location to geographical coordinates, which then maps to specific timezones).

I draw two key observations from Figure 8. First, commits are highest during office hours. Office hours are loosely characterized as 9 am to 5 pm for convenience, although this varies by region, company, and worker. Second, compared to universal time, commits based on local time are more pronouncedly bimodal. In particular, commits peak at two different timings. Once right before the standard lunch hours and once right at the end of a workday. The lunchtime dip squares with developers rarely skipping meals to be productive (Stack Exchange 2018).

I extend the analysis in Figure 8 to different days of the week. Panel (a) of Figure 9 shows that commits follow the average time-of-day cadence for all workdays (Mon–Fri), are much more likely to peak just before dinner time on Saturdays, and much flatter on Sundays. Pull requests in panel (b) of Figure 9 are also bimodal, but the second peak occurs approximately three hours before the end of the day on workdays. Otherwise, the cadence of pull requests is flat during the day hours of weekends.[9]

[9] One conjecture for why the second peak of pull requests occurs hours before the end of the day is that such timing allows team members in the same timezone to review the pull request before the end of the day. It is also generally bad practice to submit a pull request at the end of a workday, leaving colleagues little time to review proposed changes.



(a) Commits

(b) Pulls requests

Figure 9: Time-of-Day and Day-of-Week

Notes—Kernel density estimate plot of commits and pull requests by time-of-day and week-of-day. Minimal smoothing applied. Solid lines are weekdays. Dashed line is Saturdays. Dashed-dotted line is Sundays. The two vertical dotted lines are the start and end time of "office hours" (9am to 5pm). The two gray shaded areas indicate the two standard meal times (noon to 2pm for lunch time; 6pm to 8pm for dinner time).

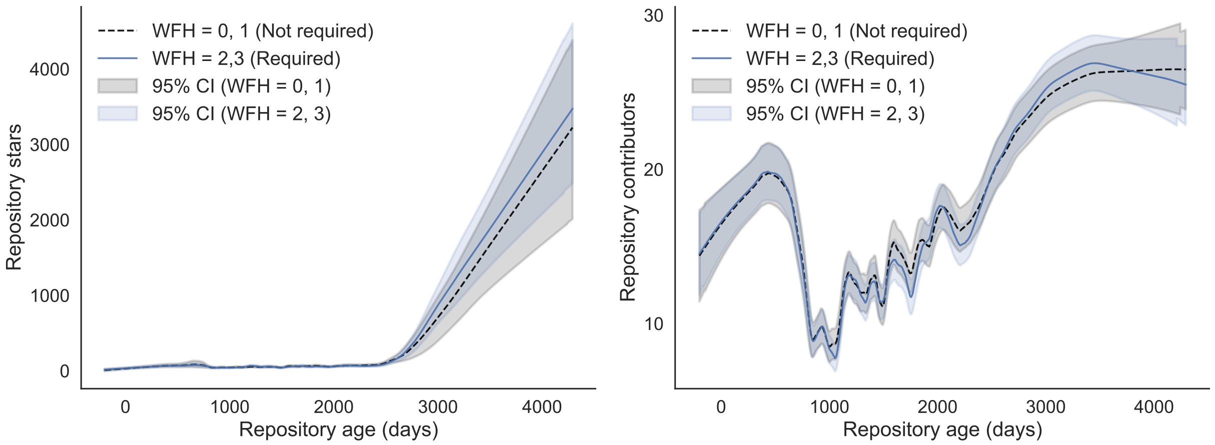
III.D User-Perceived Quality of Projects

One concern with comparing output before and after state-imposed WFH, and more generally, before and after the COVID-19 pandemic, is the quality of the work.

Similar to the problem of measuring productivity, approaching the quality of output is non-trivial. One popular approximation for repository quality on the GitHub platform is to use user-perceived code quality using stars (Papamichail et al. 2016; Sanatinia and Noubir 2016). When users like a project on GitHub and want to bookmark it for their use, they can “star” a repository, indicating project quality.

While Figure A4 suggests that there are systematic differences in stars for projects active before and after WFH, this does not account for the evolution of a project’s popularity over time. Figure 10 shows this evolution over the lifespan of projects. If anything, projects active during WFH has better quality than those before WFH, although this difference is not substantial. As a similar exercise, using the number of contributors to a project does not suggest a substantial difference in project size before and after WFH.[10]

[10] Figures A17 to A19 in the Online Appendix show that separating the recommended WFH period (WFH=1) does not change the findings. Alternative metrics of user-perceived project quality, like the number of forks, do not change the conclusion as well (Figure A20 and Figure A21).



(a) Stars

(b) Contributors

Figure 10: Metrics of Quality and Size

Notes— Blue solid line is repositories active during required WFH period. Black dashed line is repositories outside of the required WFH period. Lines are from a locally weighted smoothing with minimal smoothing. Repository age is defined by the repository create date relative to Jan 1, 2020. Shaded area indicates bootstrapped 95% confidence interval ($n = 1,000$). Figures A17 to A19 in the Online Appendix suggest similar patterns for the pull request sample and when the recommended WFH (WFH = 1) is separated.

Broadly, the descriptions from Figures 3 to 10 provide a sense of the composition of users active on the GitHub. Further, they provide some sense of how closely the cadence of tracked changes follows regular office work cadence and that the quality of projects active before and after the pandemic is similar.

IV Results

To examine if the state-imposed WFH affects output, I bin the tracked changes into individual-WFH arm cells (Bertrand et al. 2004) and estimate:[11]

$$\text{(1)} \qquad \ln(1 + \text{tracked changes})_{ik} = \alpha_i + \sum_{k \in \{0,1,(2,3)\}} \gamma_k \mathbb{1}\{\text{WFH} = k\}_i + \varepsilon_{ik},$$

where the outcome is the log of commits, and separately, pull requests, per user i per day in the WFH arm = k period.

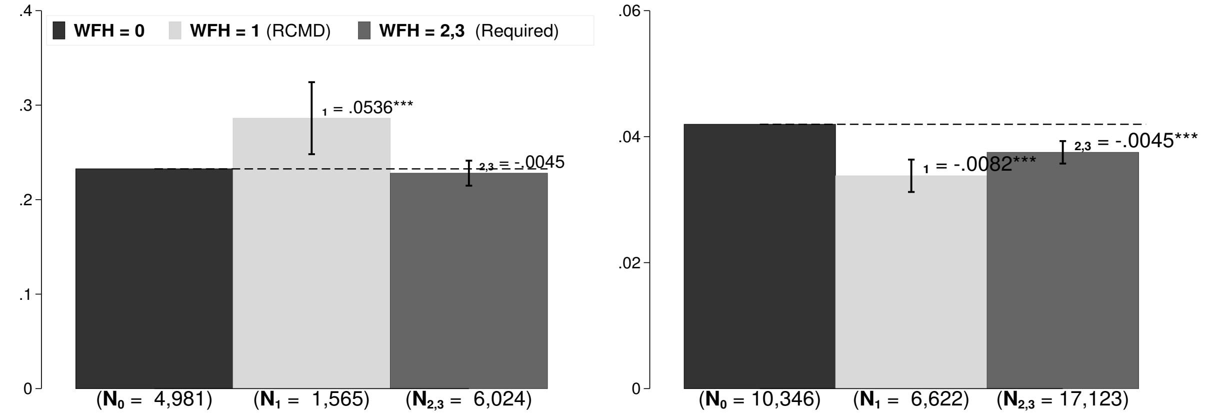
[11] This specification does not use the variation across region and time since the tracked changes are relatively infrequent at the user, project, and day (or even weeks and months) level. However, the event study analyses fully reported in the Online Appendix use the variation across regions and days (see Footnote 12). The findings are broadly similar. While there are distinct changes in the observed activity after day 0 of lockdowns, and while there is a slight trend in the observed effect, the change in output following lockdowns is minimal once we control for the profile of the active users.

I collapse the OxCGRT WFH coding (Table 1) for 2 and 3 into a single dummy that captures binding state-imposed WFH. α_i is the individuals fixed effects to account for user-specific cadences in software delivery since different GitHub users will have different priors about how many bundled changes should happen before they commit resources to archive those changes. In software engineering terms, one key assumption regarding the individual fixed effects is that the commits and pull requests of individuals, when scaled by the size of code changes, are the same before and after lockdowns (see Jaspan and Sadowski 2019).

Subject to the identifying assumptions, the γ_k estimates capture the impact of state-imposed WFH on the individual work pattern of developers. The key identifying assumption is that the precise timing of state-imposed WFH, for regions that eventually imposed WFH, is independent of the dynamics in output. In the specific context of this study, the state-imposed WFH timing should be uncorrelated with individual-specific cadences in software delivery. The panel is balanced in that it includes only users active during the non-WFH period (WFH = 0) and active during at least one of the recommended WFH (WFH = 1) or required WFH (WFH = 2, 3) periods. Standard errors are clustered by the regions to allow for correlation in work patterns between individuals residing in the same region.[12]

Figure 11 reports the results from estimating Equation (1) for commits and pull requests separately. Panel (a) suggests that state-imposed WFH has no negative impact on commits at the user level. The estimate of -.0045 for $\gamma_{2,3}$ implies that commits per user-day decrease by only approximately .5 percent, but this is not statistically significant. For recommended WFH, the estimate is positive and implies a 5.4 percent increase in output ($p < 0.01$).

[12] An empirical test of the assumption that state-imposed WFH timing is uncorrelated with output dynamics is to test how output evolves leading up to the start of WFH. While this test is not readily an option because tracked changes for any given individual and project are not at a sufficiently high frequency, Appendix D in the Online Appendix shows no pre-trends in the output metrics at the region-day, using an event study specification with a 21–day window before and after the start of WFH. These exercises in the Online Appendix are cognizant of, but do not dive fully into, the emerging difference-in-differences literature on the variation of treatment timing across groups of units in the sample (Baker et al. 2021; Callaway and Sant'Anna 2020; Goodman-Bacon 2019).



(a) Commits

(b) Pull requests

Figure 11: WFH Impact on Tracked Changes (User Level)

Notes—Figure plots the estimated impact (estimates of γ_k from Equation (1)) of state-imposed WFH at the user-day level. The dependent variables are commits and pull requests per individual per day in a WFH arm. The first bar in each subfigure indicates the baseline—WFH=0 (no WFH). Subsequent bars add back the estimated impacts to the baseline estimate ($\gamma_0 + \gamma_\ell$, $\ell = 1$ or $2, 3$). Annotated estimates in figures are the estimates of γ_k. ***, **, and * denotes significance at the 1, 5, and 10 percent level, respectively. Parenthesized numbers ($\mathbf{N}_k$) below bars indicate size of the individual observations for the corresponding WFH arm. Capped vertical bars are 95% confidence intervals from robust standard errors clustered by country.

Panel (b) of Figure 11 reports the estimates for pull requests as the metric of output. The estimate of -.0045 for $\gamma_{2,3}$ implies that pull requests per user-day decrease by approximately .5 percent ($p < 0.01$). The decrease in pull requests is larger during the recommended WFH period, with the estimate of -.0082 implying that pull requests per user-repository-day fell by .8 percent ($p < 0.01$).

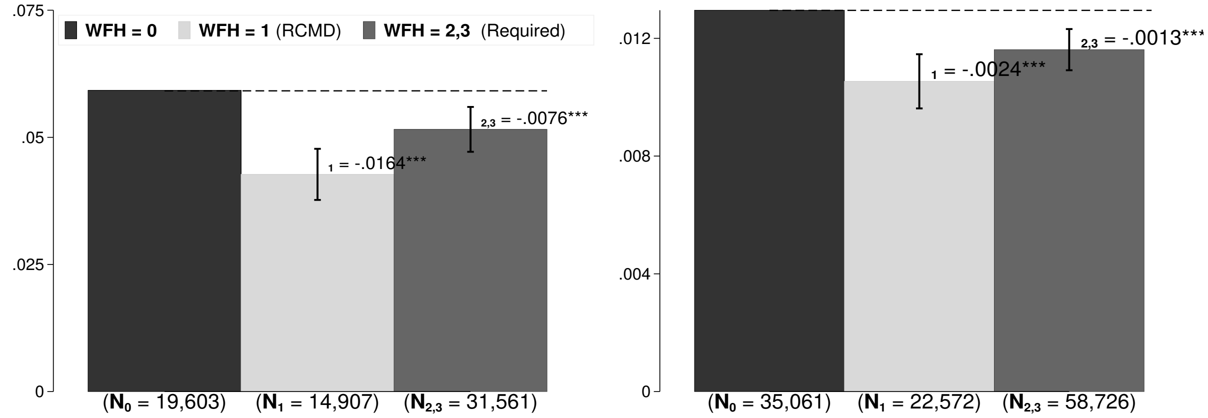
Figure 12 shows the results when the unit of analysis is defined at the user-repository-WFH arms. This approach, with repository fixed effects, allow for project-specific cadences. The findings are similar, except that the estimated impact of WFH for commits is now negative and statistically significant. However, the conclusion of a limited WFH impact on output still holds. Since the estimates are precise, one can rule out even modest gains in output (confirmed by one-sided t-tests).[13]

[13] The coefficients from Figure 12 are from

$$\ln(1 + \text{tracked changes})_{ijk} = \alpha_i + \alpha_j + \sum_{k \in \{0,1,(2,3)\}} \gamma_k \mathbb{1}\{\text{WFH} = k\}_i + \varepsilon_{ijk}, \tag{2}$$

where the outcome is log commits or log pull requests per user i in repository j per day in the WFH arm = k period. α_i and α_j are the user and repository fixed effects to account for user- and project-specific cadences in software delivery. γ_k estimates capture the impact of state-imposed WFH on the individual work pattern of developers. Standard errors are clustered by the countries.

The repository fixed effects subsume programming language fixed effects (see Table A1 and Table A2 in the Online Appendix for the distributions of languages). Projects with different languages



(a) Commits

(b) Pull requests

Figure 12: WFH Impact on Tracked Changes (User-Repository Level)

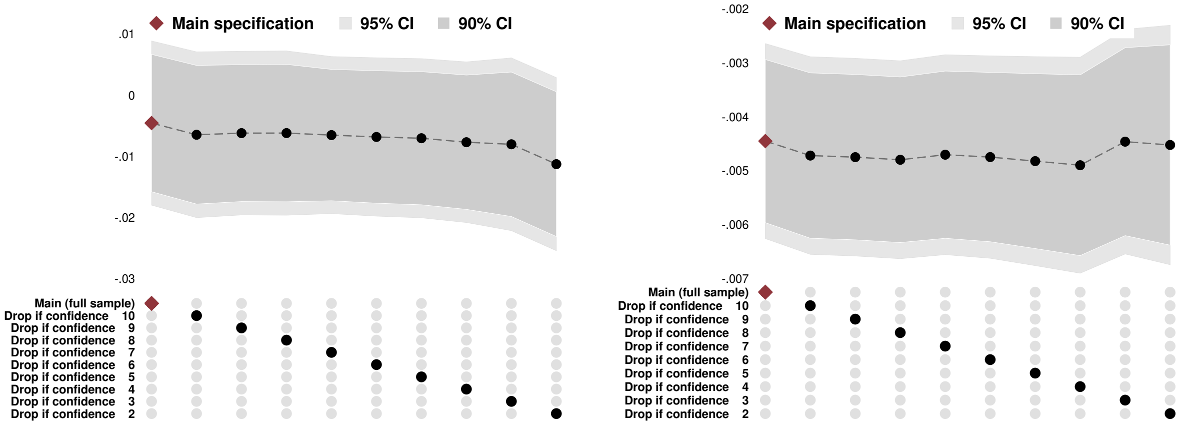
Notes—Figure plots the estimated impact (estimates of γ_k from Equation (2)) of state-imposed WFH. The dependent variables are commits and pull requests per individual-repository per day in a WFH arm. The first bar in each subfigure indicates the baseline—WFH=0 (no WFH). Subsequent bars add back the estimated impacts to the baseline estimate ($\gamma_0+\gamma_\ell$, $\ell = 1$ or $2, 3$). Annotated estimates in figures are the estimates of γ_k. ***, **, and * denotes significance at the 1, 5, and 10 percent level, respectively. Parenthesized numbers ($\mathbf{N}_k$) below bars indicate size of the individual-repository observations for the corresponding WFH arm. Capped vertical bars are 95% confidence intervals from robust standard errors clustered by country.

One concern is that any changes in output we observe are an artifact of more or fewer work duties because more bugs are being discovered due to more users on the GitHub platform after lockdowns. In Appendix D of the Online Appendix, I show using event studies with a 42–days window that there are no discernible changes in the opening and closing of issues (which include bug reporting) after lockdowns. This should mitigate concerns that changes observed in output arise because of an increase or decrease in software bug discovery.

Figure 13 shows that the main estimates are insensitive to the quality of the geocoding of user self-reported location to regions. I iteratively drop users, starting with the worse geocoding quality, from the sample and re-estimate Equation (2) where the estimates do not vary dramatically.

A different question of interest is how well the tracked changes on GitHub approximate the output of employees of firms. This affects the interpretation of what type of activity is changing during lockdowns. In Section II.C above, I cite developer-centric surveys reporting how coding purely as a hobby instead of work is very rare.

(or frameworks) might have different sizes. For example, C++ projects are larger and more complicated than Java projects at Baidu. In addition, certain (C++) projects require connection to more powerful remote machines for unit testing, which induces additional logistical complications when WFH (see Bao et al. 2020).



(a) Commits

(b) Pulls requests

Figure 13: Sensitivity to Geocoding

Notes—Plot shows how the required WFH estimates ($\gamma_{2,3}$), from estimating Equation (2) and as reported in the micro-sample results in Figure 12, changes when activities from users that are less confidently matched to regions are progressively dropped from estimation. The "Confidence" number, as returned by the OSM API, *increases* as the API is *less* certain about the geocoded region. This is also corroborated by a random sample of a 1,000 geocodings, which I check by hand (see Appendix E in the Online Appendix). The first marker in red diamond is the main result as reported in Figure 11.

In Section III, the descriptive results suggest that tracked changes follow the regular workdays and office hours cadence. Here, I address the concern analytically. The metadata from the archive does not directly contain information on whether commits are from employees or non-employees. However, the metadata includes the organization that individuals work at, should they choose to list it. I repeat the analyses for only those individuals who list their organizations, with the estimates reported in Figure A5.

Additionally, I build a new panel that includes only tracked changes during office hours on workdays (Mon–Fri). The office hours are between 8 am to 6 pm local time as inferred by the local timezone. This is one more hour before and one more hour after the 9 am to 5 pm "office hours". The longer hours accommodates findings from the literature about longer work hours during the pandemic (e.g. DeFilippis et al. 2022; Forsgren 2020; Friedman 2020; McDermott and Hansen 2021). Using this new panel of tracked changes that occur only during workday office hours, I re-analyze how output changes after lockdowns. The estimates are reported in Figure A6.

The two approaches need not necessarily capture commits and pull requests as strictly company-required work activities. But they get us closer to the concept of work activity as opposed to hobby projects (as discussed in Section II.C). In both cases, using only (i) the sample of users with organizations in Figure A5 and (ii) using only tracked changes occurring during workday office hours in Figure A6, the estimates remain statistically negative and are the same magnitude as the main estimates.

V Discussion

V.A Non-compliance With WFH Policies

Ideally, one can observe whether individuals in the sample work from home or office. Such information lets us know which individuals comply with the mandatory WFH policy. In reality, compliance with WFH policies is unobserved. All estimates, therefore, fall under an intention-to-treat basis (ITT, Angrist et al. 1996) since we observe assignment to state-imposed WFH but not compliance.

The ITT estimates show how output changed after the state-imposed mandatory WFH policy. This is the most policy-relevant question. From the results Figure 11 in Section IV, there is little detected effect (< 0.5 percent with a standard error of 0.091 percent) of mandated WFH on output. This estimate, however, encompasses both the types of individuals who do and do not comply with the WFH mandates. The proportion of compliers (compliance rate) and how the non-compliers are affected by WFH affect the estimate of WFH on people who switched to WFH once the lockdown starts.

The discussion below, therefore, addresses three points: (i) the sources of non-compliance, (ii) why the individuals already WFH before the pandemic are the main source of non-compliance, and (iii) how the effect of WFH on individual output should still be small even high levels of non-compliance.

In the ITT framework, compliers are individuals who comply with the mandatory WFH policies by working from home during WFH periods and working from office during non-WFH periods. These are the individuals of interest if one is concerned about how WFH affected output. Compliers, however, are not the only possible type of individuals in this policy setting. The ITT framework classifies three groups of non-compliers: defiers, never-takers, and always-takers. The proportion of these non-compliers will affect the estimate of how WFH affects the output of individuals.

First, defiers are individuals who WFH pre-pandemic and return to work from office during the WFH periods while everyone else in their region is working from home. I rule out individuals with this peculiar behavior. The second group of non-compliers, the never-takers, are individuals who never WFH before or after the pandemic. Given that the sample of individuals is GitHub users who mostly work in ICT, I find it reasonable to rule out such never-takers, who never WFH even after lockdowns have started. If anything, never-takers more accurately describe workers in the essential services.[14]

The final group of non-compliers is the always-takers, who have been working from home since before the pandemic began. Since tasks such as writing code can be easily done remotely, it is not reasonable to rule such individuals out. Hence, one needs to consider how large the always-takers are in the sample. The larger the proportion of individuals in the sample who have always been working from home since before the pandemic, the smaller the compliance rate.

Figure 14 illustrates how the compliance rate affects the back-of-envelope estimate of how the WFH policy affected the output of compliers. The smaller the compliance rate, the larger the implied effect WFH has on their output. Since actual compliance rates are unobserved, I turn to survey results to understand what compliance rates are reasonable to assume.

[14] Ruling out of defiers is the monotonicity assumption. So that under the ITT framework (Angrist et al. 1996), the $\pi_D \cdot \text{ITT}_D$ for defiers drops out from the proportion-weighted average of the ITT effects by group: ITT $= \sum_{g \in G} \pi_g \cdot \text{ITT}_g$, with $G \in \{C, D, AT, NT\}$ for compliers, defiers, always-takers, and never-takers. Ruling out never-takers further eliminates $\pi_{NT} \cdot \text{ITT}_{NT}$ from the proportion-weighted average. This means we are left with: ITT $= \pi_C \cdot \text{ITT}_C + \pi_{AT} \cdot \text{ITT}_{AT}$.

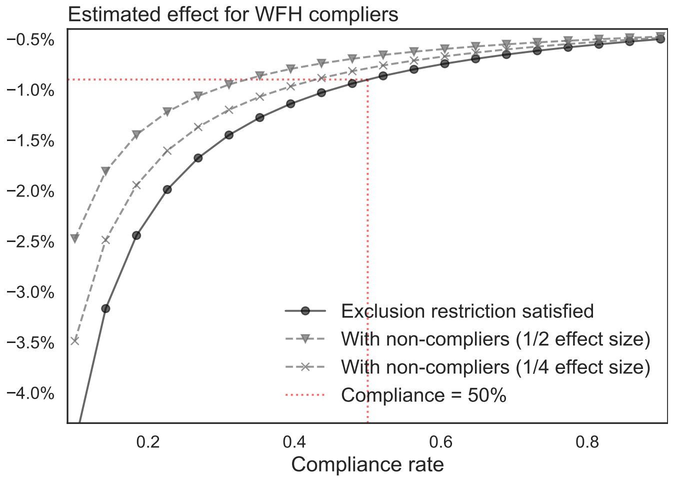


Figure 14: Effect Size Depends on Compliance

Notes—Lines are estimated WFH effect size for pull requests for individuals who comply with state-imposed WFH. Solid line is estimated effect when the exclusion restriction is satisfied $\left(\frac{\hat{\gamma}_{2,3}}{\pi_c}\right)$. Dashed line with triangle is estimated effect when the effect from the always-takers is half the estimated ITT $\left(\frac{\hat{\gamma}_{2,3}}{\pi_c} - \frac{\hat{\gamma}_{2,3}}{2}\frac{\pi_{AT}}{\pi_C}\right)$. Dashed line with cross is estimated effect when the effect from the always-takers is a quarter of the estimated ITT $\left(\frac{\hat{\gamma}_{2,3}}{\pi_c} - \frac{\hat{\gamma}_{2,3}}{4}\frac{\pi_{AT}}{\pi_C}\right)$. All lines assume no defiers and never-takers, so the only non-compliers are the always-takers. $\hat{\gamma}_{2,3} = -0.45\%$ is from panel (b) of Figure 11.

According to the Stack Exchange (2015) developer-centric survey, 29 percent of developers work at least partially remotely. The more recent Stack Exchange (2019) survey indicates that 18 percent work at least partially remotely. Another study by Yang et al. (2022) who uses employees from Microsoft reports similar numbers—approximately 18 percent of workers WFH before the pandemic.[15] [16]

Having 29 percent of developers who WFH at least half the time before the pandemic implies that the compliance rate with WFH policy is 71 percent in the absence of defiers and never-takers. To help bound the estimated effect, I assume a 50 percent compliance rate. This is lower than the compliance rate implied by the surveys (Stack Exchange 2015, 2019; Yang et al. 2022). With a 50 percent compliance rate, the ITT estimate of 0.45 percent for pull requests suggests that the negative impact of WFH on output is still less than a full one percent (the red dotted line in Figure 14).

Moreover, the implied negative impact of less than one percent is overstated if

[15] The 29 percent in Stack Exchange (2015) is derived from 10.4 percent "full-time remote" plus 18.6 percent "part-time remote". 48.1 percent "rarely work remote" and 22.8 percent "never". The 18 percent in Stack Exchange (2019) is derived from 12 percent "full-time remote" plus 3.4 percent "more than half" plus 2.8 percent "about half the time".

[16] The Stack Overflow annual developer surveys provide less conservative estimates of pre-pandemic WFH rates than the broader studies and surveys. Bartik et al. (2020) find that the share of employees who can WFH in normal times is low. Bloom (2020) reports that the share of working days spent WFH is low at 5 percent of working days in a year. Alipour et al. (2020) likewise describes low WFH rates: 3.5 percent in Germany, 5.1 percent in the U.K., and 4 percent in the U.S. from their calculations and two cited studies (Watson 2020; Mas and Pallais 2017). These numbers are low because of the broader occupations considered and because they do not distinguish between entirely WFH and only partially WFH.

the WFH policies also affected those individuals who were already working from home before the pandemic. The closure of workplaces, for instance, will coincide with closures of schools and recreational points of interest that might negatively affect work cadence. In this case, the implied negative impact is even lower (the dashed lines in Figure 14). Overall, even if there is low compliance and if the WFH policies negatively affected individuals who were already working from home, the implied negative effect of WFH on output is still smaller than a full one percent.[17]

V.B Key Related Studies

I place my findings in context with key related studies on WFH productivity and productive output.

First, the finding in this study complements the Bloom et al. (2015) seminal study on the causal effect of WFH on productivity using a randomized controlled trial in Ctrip, a travel agency in China. The study's 249 participants (before attrition) are call center representatives whose work essentially involves answering calls from customers, taking orders, and making calls to hotels and airlines to place the orders. Ensuring that employees in both the WFH and control group have the same IT equipment and internet access, Bloom et al. (2015) find that WFH improves productivity. Specifically, they find increases along both the extensive and intensive margin. The 13 percent increase in productivity comes from a 9 percent increase in working time and a 4 percent increase in calls per minute.

Using observational data from individuals in tech-related industries comple-

[17] The WFH policies also affecting individuals already working from home violate the exclusion restriction assumption. The bias is of the form $\left(-\text{ITT}_{AT} \cdot \frac{\pi_{AT}}{\pi_C}\right)$, where C refers to individuals complying with WFH policies, AT refers to individuals who were already working from home before the pandemic, and ITT_{AT} is the effect that the WFH policies have on the AT individuals.

There are a few plausible reasons why the exclusion restriction fails in this context. First, the state-imposed WFH is essentially a government response to the pandemic. This coincides with other fiscal responses that might impact both employee and employer behavior, including work arrangements and work cadence. Second, the closure of workplaces also coincides with the closure of other places (e.g., parks and recreation), which also potentially alters work patterns. Third, closure of workplaces also applies to cohabitants, including kids, during closures of schools, which potentially affects work output. Each of these concerns applies to different subgroups of the population.

ments the Bloom et al. (2015) study. The type of work captured in Bloom et al. (2015) is well-defined with obvious metrics for productivity. Since short-run outputs in tech-related (or even science-related) work have no obvious milestones (e.g., how many bugs will appear and how many features to add are seldom clear from the onset of a project), this study provides insight into how output is affected when work is non-transactional and non-routine.

The second set of studies I highlight, which surfaced during the pandemic, can be broken down into two strands. One looks at production from software developers (e.g., Bao et al. 2020; Ford et al. 2021; Forsgren 2020; McDermott and Hansen 2021) because of the tracked changes in a pipeline. The other, more broadly, is in the ICT sector (e.g., Gibbs et al. 2021; Yang et al. 2022) because work machines have software that monitors work activity.

The GitHub team in Forsgren (2020) is the first to my knowledge that uses the GitHub tracked activity data to study how the pandemic broadly affected output. Their analysis of trends suggests minimal impact to GitHub activity but a change to daily work patterns. In particular, Forsgren (2020) finds that developers are now working longer hours, a finding corroborated by McDermott and Hansen (2021). McDermott and Hansen (2021) focus on key metropolitan areas and, using past years as a counterfactual, find more work reallocated outside of traditional office hours. These two studies use the same type of activity to approximate productivity, but with very different approaches and focus than this study. Their findings about longer workdays are consistent with other studies in the literature (DeFilippis et al. 2022; Friedman 2020).

The studies by Bao et al. (2020) and Ford et al. (2021), on the other hand, focus on the productivity of software developers in one firm. Bao et al. 2020 use data from Baidu. Ford et al. 2021 use data from Microsoft. Both have descriptive estimates, and, similar to this study, both find minimal WFH impact on productivity.

More broadly, Gibbs et al. 2021; Yang et al. 2022 study WFH impact on ICT workers using a large Asian IT firm and Microsoft, respectively. The study by Gibbs et al. 2021 uses data from software installed on work machines that track employee activity and performance. They have input and output measures and can thus study productivity in the canonical definition. They find that productive activity is stable and that longer working hours are what drives observed falls in productivity from WFH. Yang et al. 2022 uses the COVID-19 pandemic as a natural experiment as well and find that collaborations among Microsoft employees suffered when WFH.

Finally, while the motivation of this study is to use real-time data as an alternative to survey-based approaches, WFH can impact intangibles (Gibbs et al. 2021). For these richer WFH concerns, a survey-based approach seems to be the way to go (e.g.,Bloom 2020; Barrero et al. 2021). Results from both can be useful complements.

V.C Limitations

This study is not without limitations, and it bears laying them out. First is the measurement of productivity. As stated in Section II.A, the canonical form of productivity is some units of output scaled by some units of input. The randomized controlled trials in Bloom et al. 2015 and Emanuel and Harrington 2021, and the studies using work trackers from companies (e.g., Ford et al. 2021; Gibbs et al. 2021), have this form. No metric for observable work time is available from GitHub. Instead, only metrics of output, such as commits and pull requests, are available and should only be interpreted as such. However, given that this study finds a small and negative WFH impact on output, and to the extent that other studies find longer working hours (e.g., Forsgren 2020; McDermott and Hansen 2021; Gibbs et al. 2021), one can still draw about changes in productivity. Still, a broad stroke conclusion that the pandemic, or WFH in general, leads to lower productivity is likely misleading since there are gains from (not) commuting (Barrero et al. 2021).

A second issue is that we only observe the state-imposed WFH timings of regions rather than of individuals. This is because GitHub does not directly provide information about whether a user is working from the office or home. This implies the results can only be interpreted as intention-to-treat effects. Section V.A discusses this in more detail. Broadly, the reliability of the estimates on individuals who comply and WFH during lockdowns will depend on compliance rate and, if there is a violation in the exclusion restriction, on the ratio of the non-compliance to compliance rate (with the bias attenuating with compliance in Figure 14).

A related issue is that since the variation comes from workplace closures during a pandemic, the findings here do not directly generalize to WFH in other scenarios. Nonetheless, to the extent that the estimates reveal the impact of WFH on output in times of adversity, given the COVID-19 pandemic, one can expect a net zero or even a positive gain in output from WFH in less arduous settings.

VI Conclusion

This study provides evidence from a large open-source and cloud-hosted software development platform on how WFH has affected individual-level output from tracked changes. While the natural experiment setting in the pandemic may not yield clean causal estimates like those in randomized experiments, this study uses output metrics and finds a negligible change during lockdowns. Output is only 0.5–0.8 percent lower when WFH. The standard errors are small, which helps rule out a dramatic impact of WFH on productivity.

Perhaps one lesson we are learning is the importance of monitoring output. Switching work environments and its effect on output may not be the biggest issue. Instead, incentivizing and monitoring staff output is the issue. The inability to monitor is why managers prefer staff to work in offices. According to the experimental setting of this study, when work cadence can be tracked objectively, it turns out that WFH has a minimal measurable impact.

As a whole, the study's findings suggest only a limited negative impact of WFH. This contributes to resolving the fundamental concern in the debate on returning to offices. Moreover, unpacking the descriptive analyses of the granular tracked changes suggests that tracked changes approximate the output of individuals. Tracked changes in a work setting are, therefore, a promising asset for big data analytics to understand both the quantity and quality of outputs.

References

Alipour, J.-V., O. Falck, and S. Schüller (2020). Germany's Capacities to Work from Home. *CESifo Working Paper No. 8227*.

Angelucci, M., M. Angrisani, D. Bennett, A. Kapteyn, and S. Schaner (2020). Remote Work and the Heterogeneous Impact of COVID-19 on Employment and Health. *IZA DP No. 13620*.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Baker, A. C., D. F. Larcker, and C. C. Y. Wang (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates?

Baltes, S. and S. Diehl (2018). Towards a Theory of Software Development Expertise. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, New York, NY, USA, pp. 187–200. Association for Computing Machinery.

Banbury, S. and D. C. Berry (1998). Disruption of office-related tasks by speech and office noise. *British Journal of Psychology 89*(3), 499–517.

Bao, L., T. Li, X. Xia, K. Zhu, H. Li, and X. Yang (2020). How does Working from Home Affect Developer Productivity? A Case Study of Baidu During COVID-19 Pandemic. pp. 1–17.

Barrero, J. M., N. Bloom, and S. J. Davis (2021). Why Working from Home Will Stick. *National Bureau of Economic Research Working Paper Series No. 28731*.

Bartik, A. W., Z. Cullen, E. L. Glaeser, M. Luca, and C. Stanton (2020). What Jobs are Being Done at Home During the COVID-19 Crisis? Evidence from Firm-Level Surveys. *HBS No. 20-138*.

Baruch, Y. (1996, jan). Self performance appraisal vs direct-manager appraisal: A case of congruence. *Journal of Managerial Psychology 11*(6), 50–65.

BBC (2020a). Coronavirus: Fujitsu Announces Permanent Work-From-Home Plan.

BBC (2020b). Coronavirus: Twitter Allows Staff to Work From Home 'Forever'.

BBC (2020c). Facebook and Google Extend Working From Home to End of Year.

BBC (2020d). Netflix Boss: Remote Working Has Negative Effects.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, feb). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bloom, N. (2020). How Working From Home Works Out. *Stanford Institute for Economic Policy Research*, 1–8.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2015). Does Working from Home Work? Evidence from a Chinese Experiment. *The Quarterly Journal of Economics 130*(1), 165–218.

Brynjolfsson, E., J. J. Horton, A. Ozimek, D. Rock, G. Sharma, and H.-Y. TuYe (2020). Covid-19 and Remote Work: an Early Look At Us Data.

Callaway, B. and P. H. C. Sant'Anna (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics*.

Choudhury, P., W. W. Koo, and X. Li (2020). Working (From Home) During a Crisis: Online Social Contributions by Workers During the Coronavirus Shock. *HBS No. 20-096*.

Choudhury, P. R., C. Foroughi, and B. Larson (2021). Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management Journal 42*(4), 655–683.

Chrystal, A. and P. Mizen (2003, jan). Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy. *Cent Bank Monet Theory Pract Essays Honour Charles Goodhart 1*.

DeFilippis, E., S. M. Impink, M. Singell, J. T. Polzer, and R. Sadun (2020). Collaborating During Coronavirus: The Impact of COVID-19 on the Nature of Work. *HBS No. 21-006*.

DeFilippis, E., S. M. Impink, M. Singell, J. T. Polzer, and R. Sadun (2022). The impact of COVID-19 on digital communication patterns. *Humanities and Social Sciences Communications 9*(1), 180.

Emanuel, N. and E. Harrington (2021). "Working" Remotely? Selection, Treatment, and the Market Provision of Remote Work. pp. 1–83.

Ford, D., M.-A. Storey, T. Zimmermann, C. Bird, S. Jaffe, C. Maddila, J. L. Butler, B. Houck, and N. Nagappan (2021, dec). A Tale of Two Cities: Software Developers Working from Home during the COVID-19 Pandemic. *ACM Trans. Softw. Eng. Methodol. 31*(2).

Forsgren, N. (2020). Octoverse Spotlight: An Analysis of Developer Productivity, Work Cadence, and Collaboration in the Early Days of COVID-19.

Friedman, A. (2020). Proof Our Work-life Balance Is in Danger (But There's Still Hope). *Atlassian*.

Gibbs, M., F. Mengel, and C. Siemroth (2021). Work From Home and Productivity: Evidence From Personnel and Analytics Data on IT Professionals. *SSRN Electronic Journal*.

GitLab (2018). 2018 Global Developer Report.

Goodhart, C. A. E. (1984). *Monetary Theory and Practice : The UK Experience*. London: Macmillan.

Goodman-Bacon, A. (2019). Difference-in-Differences With Variation in Treatment Timing. *Working paper*.

Gottlieb, C., J. Grobovsek, and M. Poschke (2020). Working From Home across Countries. *CEPR*.

Jaspan, C. and C. Sadowski (2019). No Single Metric Captures Productivity BT - Rethinking Productivity in Software Engineering. pp. 13–20. Berkeley, CA: Apress.

Khanna, V. (2020). The Economic Consequences of Working From Home. *The Straits Times Covid 19 Special: Economic Affairs*.

Kummer, M., O. Slivko, and X. M. Zhang (2020, aug). Unemployment and Digital Public Goods Contribution. *Information Systems Research 31*(3), 801–819.

Lerman, R. and J. Greene (2020). Big Tech Was First to Send Workers Home. Now It's in No Rush to Bring Them Back. *The Washington Post*.

Mas, A. and A. Pallais (2017). Valuing Alternative Work Arrangements. *American Economic Review 107*(12), 3722–3759.

McDermott, G. R. and B. Hansen (2021). Labor Reallocation and Remote Work During COVID-19: Real-time Evidence from GitHub. *National Bureau of Economic Research Working Paper Series No. 29598*.

Ozimek, A. (2020). The Future of Remote Working.

Papamichail, M., T. Diamantopoulos, and A. Symeonidis (2016). User-Perceived Source Code Quality Estimation Based on Static Analysis Metrics. In *2016 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pp. 100–107.

Petherick, A., T. Hale, T. Phillips, and S. Webster (2020). Variation in Government Responses to COVID-19. *Blavatnik School Working Paper*.

Ralph, P., S. Baltes, G. Adisaputri, R. Torkar, V. Kovalenko, M. Kalinowski, N. Novielli, S. Yoo, X. Devroey, X. Tan, M. Zhou, B. Turhan, R. Hoda, H. Hata, G. Robles, A. Milani Fard, and R. Alkadhi (2020). Pandemic programming. *Empirical Software Engineering 25*(6), 4927–4961.

Ruprechter, T., M. H. Ribeiro, T. Santos, F. Lemmerich, M. Strohmaier, R. West, and D. Helic (2021). Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions. *arXiv*, 1–31.

Sanatinia, A. and G. Noubir (2016). On GitHub's Programming Languages. *ArXiv abs/1603.0*.

Spataro, J. (2020). Helping Our Developers Stay Productive While Working Remotely.

Stack Exchange (2015). Stack Overflow developer survey 2015.

Stack Exchange (2018). Stack Overflow developer survey 2018.

Stack Exchange (2019). Stack Overflow developer survey 2019.

Stack Exchange (2022). Stack Overflow developer survey 2022.

Stanton, C. T. and P. Tiwari (2021). Housing Consumption and the Cost of Remote Work. *NBER Working Paper 28483*.

Uddin, G., O. Alam, and A. Serebrenik (2022). A qualitative study of developers' discussions of their problems and joys during the early COVID-19 months. *Empirical Software Engineering 27*(5), 117.

Watson, B. (2020). Coronavirus and Homeworking in the UK Labour Market: 2019. Technical report.

Whiting, K. (2020). Is Flexible Working Here to Stay? We Asked 6 Companies How to Make It Work. *World Economic Forum*.

Yang, L., D. Holtz, S. Jaffe, S. Suri, S. Sinha, J. Weston, C. Joyce, N. Shah, K. Sherman, B. Hecht, and J. Teevan (2022). The effects of remote work on collaboration among information workers. *Nature Human Behaviour 6*(1), 43–54.

YouGov (2020). Huffpost: Working From Home.

Zlotnick, F. (2017, June). GitHub Open Source Survey 2017. `http://opensourcesurvey.org/2017/`.

A Appendix

A.1 Data build details

To build the user-repository-WFH panels, I proceed as follows:

1. From GitHub's open-access public dataset on Google BigQuery, I query user commits from `Jan-Jun 2020`. This gives a user-commit log record with timestamps, author names, and repository name.

2. I curate a list of usernames in four ways: (a) using the author name of a commit through the GitHub search API, (b) using the username string from the repository name (e.g. `johnx/projecta` implies `projecta` belongs to user `johnx`), and later from (c) users who raise/close issues, and (d) users who submit a pull request. To retrieve user-level information (e.g. location, user type, repositories, etc.) from the list of usernames, I query GitHub's User API. The majority of usernames are successfully queried (620,922 of 626,488 or 99.1%), with the minority having 404 or 502 HTTP response error codes at the time of query.

3. *Geocoding.* To geocode the users from GitHub to a country (or U.S. state), I query the `OpenStreetMap (Nominatim) API` from the Python `geocoder` library, using the location strings that users enter into their account. Approximately half the users have a location string (309,247 or 49.4%). Each successful query returns a hierarchy of geographical information `country-state-city-county`, with a confidence score for reliability of the result (see Table Appendix E). For geocoding to country-dates to get the government-enforced WFH status, I retain only country for non-U.S. countries and states for the U.S. observations (because this is the level of granularity in the OxCGRT, see below).

   Almost all users with a location string can be geocoded to a country/U.S. state (303,403 or 98.1%). Almost all location strings can be geocoded—there are 47,681 unique location strings from all the users in the sample, of which 42,552 (89.24%) can be successfully geocoded.

4. *WFH records.* From the geocoding, each activity record (commits, raise/closing of issues, pull requests) have a (`country/U.S. state - date`) tuple that I can then map to government enforced-WFH (work from home/closure of workplaces) policies in the OxCGRT. This is their `C2_Workplace` indicator. Cognizant that federal and state governments may vary in their timing, I retain the flag indicators for when a policy was targeted at sub-regions instead of a nation-wide enforcement (for subsequent robustness checks). This is their `C2_Flag`.

   In the panels, for all non-U.S. countries, the WFH indicator for each user-repo-WFH cell is the WFH based on the user's country. For those users geocoded to the U.S. states, the WFH indicator is based on the individual U.S. states, since this is the level of granularity that the OxCGRT currently offers.

Alternatively, for county-level policies of the U.S. sample, I combine two sources of business closure records: a complete record at the state level from the COVID-19 US State Policy Database[18] and a partial record of 569 counties from crowdsourcing[19]. Where available, I use the [business_closed_date] record from the county-level crowdsourced records. For the remaining counties, I use their state-level [Closed other non-essential businesses] record. From these, 258 counties have earlier localized closures relative to the state, while 37 counties have later closures.

5. *Repository records.* Records directly from the commits archive contains 245,506 repositories. To get their repo-level information (e.g. contributors, language, open issues, etc.), I use the GitHub Repository API. Most of the repositories can be succesfully queried at the time of query (240,150 or 97.9%).

   A minority of commits in the sample are to multiple repositories (usually the same project under different forks). I map these multi-repository commits to the original repository (forked = False), so that language, contributors, and open issues records etc. belong to the original repository.

6. *Pull requests and issues record.* From the initial repository records, I also query their historical record of pull requests and issues (both opening and closing), retaining only records created in the year 2020. These records, with the user record, are mapped to a WFH status as described above. For closure of issues, I use the recorded assignee as the user who "resolved the issue". This step yields 39,958 closed issues that can be geocoded (included in analyses); 253,632 opened issues, and 288,175 pull requests.

7. *Standard to local timezone inference and conversion.* The GitHub archive of commits come with Unix timestamps. These are the number of seconds elapsed since Unix epoch on 00:00:00 UTC on 1 January 1970. To convert the commits from standard time to local time, I first convert the Unix timestamps back to UTC (Coordinated Universal Time) with zero offset, where UTC is now the worldwide standard and bypasses problems and confusion about daylight saving time.

   Then, to convert the timzone-agonstic timestamps to local time based on the user-reported location, I first query OpenStreetMaps to retrieve longitude and latitude coordinates. I then lookup the timezone for the given set of geo-coordinates. This step relies on the Olson database where timezones are represented by polygons and timezone membership is based on the longitude-latitude coordinate. See Figure A9 in Appendix B of the Online Appendix for details.

   Finally, I convert the timezone-agnostic timestamps to local time based on

[18] http://www.tinyurl.com/statepolicysources.

[19] https://docs.google.com/spreadsheets/d/133Lry-k80-BfdPXhlS0VHsLEUQh5_UutqAt7czZd7ek/edit#gid=0.

the inferred local timezone. All analyses involving timestamps are in 24-hour local time.

A.2 Figures and Tables

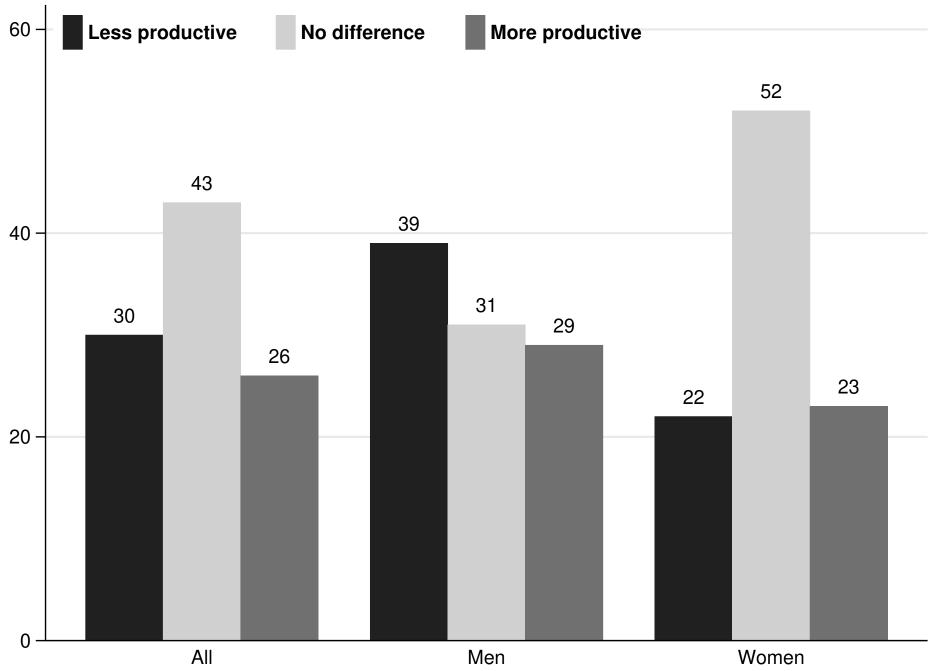


Figure A1: Self-Reported Impact of WFH on Productivity

Notes—Plot shows YouGov 2020 survey responses of 1,000 US respondents, conducted May 2020, on whether WFH improves productivity. Y-axis is percentage. Numbers do not add up to 100 percent because those who responded "not sure" are not included.
Source: YouGov (2020).

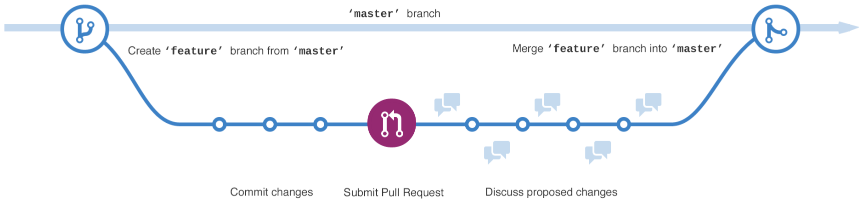

'master' branch
Create 'feature' branch from 'master'
Merge 'feature' branch into 'master'
Commit changes
Submit Pull Request
Discuss proposed changes

Figure A2: Git Branching Workflow

Notes—Figure shows the basic Git workflow, focusing on commits and pull requests in the context of a main deployable branch. Graphic taken directly from source.
Source: https://guides.github.com.

Table A1—*List of Countries That Rollback WFH in Sample*

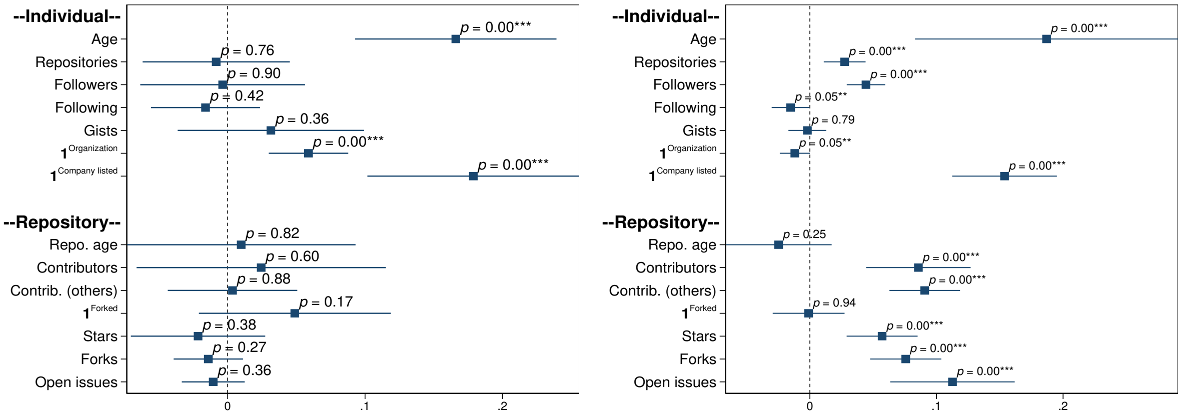
Country	WFH Day 0	Rollback Day 0	WFH rolled back to
Turkmenistan	24Mar	1Apr	0
Ghana	30Mar	20Apr	1
Greenland	18Mar	27Apr	0
Cameroon	18Mar	1May	0
Italy	22Feb	4May	1
Greece	12Mar	5May	0
Lesotho	18Mar	6May	1
Mali	25Mar	10May	1
Luxembourg	16Mar	11May	1
Burkina Faso	21Mar	14May	0
Australia	23Mar	15May	1
Slovenia	20Mar	18May	0
Botswana	2Apr	20May	1
Czech Republic	13Mar	25May	1
Timor-Leste	28Mar	27May	0
Bangladesh	19Mar	31May	1
Laos	30Mar	1Jun	0
Rwanda	21Mar	2Jun	0
Slovak Republic	13Mar	3Jun	0
Thailand	17Mar	6Jun	1
Tunisia	22Mar	8Jun	0
Mauritius	20Mar	12Jun	0
Guinea	27Mar	15Jun	1
Falkland Islands	26Mar	16Jun	0
Romania	12Mar	17Jun	1
Singapore	7Apr	19Jun	1
Saudi Arabia	16Mar	21Jun	1
Cayman Islands	23Mar	22Jun	0
Morocco	20Mar	24Jun	1
Ireland	27Mar	26Jun	1
Dominica	1Apr	27Jun	0

Notes—Table enumerates countries that rolled back state-imposed WFH during the sample period of Jan–Jun 2020, corresponding to timing of rollbacks in Figure 2. Second column shows the first day of state-imposed WFH (OxCGRT WFH $\in$ {2, 3}). Third column shows the first day of rolling back state-impose WFH; that is, having the WFH indicating step down from either 2 or 3 to a 0 or 1.

Table A2—*Selected Treated Countries as Reference*

Country	Day 0	Commits	Users
Portugal	12Mar	2,348	676
Norway	12Mar	2,179	733
Romania	12Mar	1,064	356
Greece	12Mar	614	253
Austria	16Mar	3,359	1,059
Turkey	16Mar	2,159	676
Hungary	16Mar	1,705	562
Chile	16Mar	1,239	278
Luxembourg	16Mar	1,015	287
Sri Lanka	16Mar	868	312
Egypt	16Mar	460	187
Lithuania	16Mar	430	82
Honduras	16Mar	138	53
France	17Mar	15,457	5,508
Russia	17Mar	8,441	3,013
Switzerland	17Mar	3,661	1,546
Ukraine	17Mar	3,173	1,023
Brazil	17Mar	3,143	1,224
Thailand	17Mar	329	165
Bosnia and Herzegovina	17Mar	173	54
Seychelles	08Apr	143	54
Japan	—	20,305	6,340
Sweden	—	4,317	1,627
Taiwan	—	1,807	668
Bulgaria	—	1,477	276
Belarus	—	1,061	367

Notes—Selected treated countries for reference with Figure A25. Countries shown here are those with WFH enforcement starting on 12 March, 16 March, 17 March, and 8 April for the year 2020. Those countries with small sample share (commits < 100) are not shown. Countries sorted by WFH enforcement date and commits size in the sample period (Jan–Jun 2020).



(a) Commits sample

(b) Pull requests sample

Figure A3: US vs. Rest of World

Notes—Difference in means for the micro-samples. Difference is between the US vs. the rest-of-world, estimated by regressing the baseline covariates on the US dummy and performing a t-test for the dummy. Tables A9–A12 Standard errors are robust. ***, **, and * denotes significance at the 1, 5, and 10 percent level, respectively.

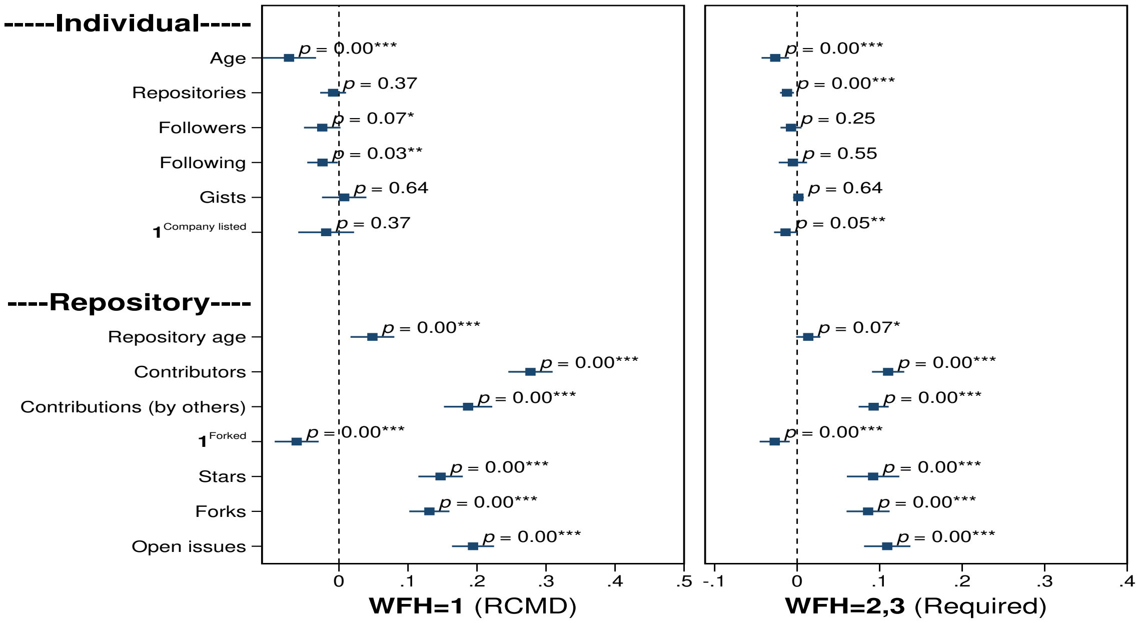
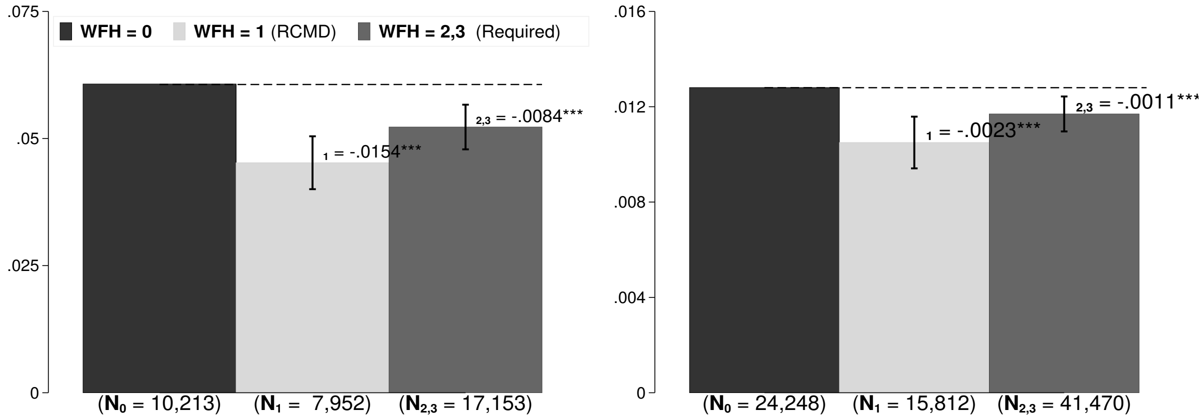


Figure A4: Differences in Observables (Commits)

Notes—Differences in means for WFH=0 (no state regulation), WFH=1 (recommended WFH), WFH=2,3 (required WFH), using the microsample from the commits records. Units in standard deviations. For the repository characteristics, repository age is defined as creation date minus 1 Jan 2020; contributions is total number of commits, pull requests, or number of issues opened; the dummy for forked indicates whether the repository was branched out from a preexisting one; stars is a measure of impact (used as a like or bookmark); forks is the number of branching out by other users; and open issues refers to the number of unresolved issues listed in the project. Tables A5–A6 of the Online Appendix tabulates the above results. Number of individuals and repositories captured are 22,183 and 25,859. The individual and repository level observations are clustered by country and programming language, respectively. ***, **, and * denotes significance at the 1, 5, and 10 percent level, respectively.

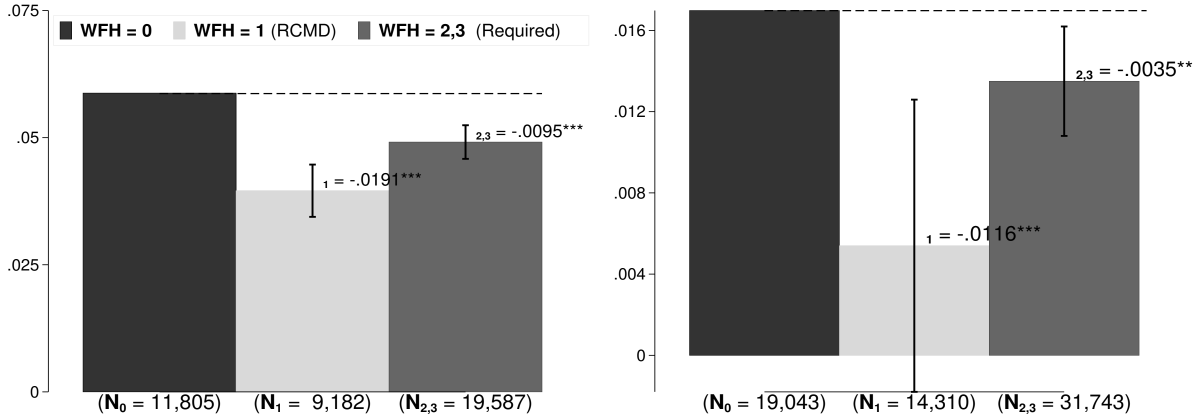


(a) Commits

(b) Pull requests

Figure A5: WFH Impact on Tracked Changes (User-Repository Level, Reported Organizations)

Notes—Figure plots the estimated impact (estimates of γ_k from Equation (2)) of state-imposed WFH. The specification is similar to Figure 12, except that only users who self-report the organization they work at are included in the estimation. Parenthesized numbers ($\mathbf{N}_k$) below bars indicate size of the individual-repository observations for the corresponding WFH arm. Capped vertical bars are 95% confidence intervals from robust standard errors clustered by country.



(a) Commits

(b) Pull requests

Figure A6: WFH Impact on Tracked Changes (User-Repository Level, Office Hours on Workdays)

Notes—Figure plots the estimated impact (estimates of γ_k from Equation (2)) of state-imposed WFH. The specification is similar to Figure 12, except that only tracked changes (commits and pull requests) during office hours and workdays are included in computing tracked changes per day. “Office hours” are between 8 am to 6 pm local time (as inferred by the timezones). Later than usual office hours accommodates findings from the literature about longer workdays after lockdowns (e.g. DeFilippis et al. 2022; Forsgren 2020; Friedman 2020; McDermott and Hansen 2021). Parenthesized numbers ($\mathbf{N}_k$) below bars indicate size of the individual-repository observations for the corresponding WFH arm. Capped vertical bars are 95% confidence intervals from robust standard errors clustered by country.