Essays in Environmental and Labor Economics

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Beate Thies

im Frühjahrs-/Sommersemester 2023

Abteilungssprecher:	Prof. Klaus Adam, PhD
Referent:	Prof. Ulrich J. Wagner, PhD
Koreferentin:	Prof. Kathrine von Graevenitz, PhD
Vorsitzender der Disputation:	Prof. Dr. Nicolas R. Ziebarth

Tag der Disputation:	07.06.2023

Acknowledgements

First and foremost, I want to thank my advisors Ulrich Wagner and Kathrine von Graevenitz for their guidance and support. I benefited greatly from their helpful comments and advice throughout the years, and especially during the time of the job market.

I thank my co-author Felix Holub for many constructive discussions and our fruitful collaboration over the last years on Chapters 1 and 3, and my co-author Ingo Isphording for our productive collaboration on Chapter 3.

This thesis also benefited from feedback and comments by former and current members of the department of economics at the University of Mannheim, especially the environmental economics group. In particular, I want to thank Nicolas Ziebarth for helpful feedback on Chapter 1 and Andreas Gerster for organizing the internal environmental economics workshop which provides an ideal setting to present early-stage projects.

I am grateful for financial support from the Graduate School of Economics and Social Science at the University of Mannheim and for support in administrative matters by Golareh Khalilpour, Ulrich Kehl, Marion Lehnert, and Caroline Mohr at the Graduate School and by Sonja Collet, Katrin Ramseier and Sylvia Rosenkranz at the Department of Economics.

I am also grateful to have met fellow students and friends who accompanied me on my journey throughout the last years. In particular, I want to thank Kinga, Linnéa, Stefan, and Vera for motivational support and for making the PhD studies a more fun and memorable time, and Johannes for a great time while we shared an office.

Most of all, I thank my parents, my brother and my sister for their continued encouragement and support.

Contents

List of Figures

List of Tables

Preface

This dissertation consists of three self-contained chapters studying questions in the fields of environmental and labor economics.

Chapter 1, co-authored with Felix Holub, studies how air quality affects productivity and work patterns among highly skilled knowledge workers. We use data from *GitHub*, the world's largest coding platform, to construct a panel including information on daily output, working hours, and task complexity for a sample of 25,000 software developers across four continents during the period 2014-2019. We combine this with information on concentrations of fine particulate matter ($PM_{2.5}$). To identify the causal effects of interest, we use an instrumental variable strategy based on wind direction and carefully control for other weather characteristics. We find that an increase in air pollution reduces output, measured by the number of total actions performed on GitHub per day, and induces developers to adapt by working on easier tasks and by ending work activity earlier. To compensate, they work more on weekends following high-pollution days, which suggests adverse impacts on their work-life-balance. The decline in output arises even at concentrations in line with current regulatory standards in the EU and US and is driven by a reduction in individual coding activity, while interactive activities are unaffected. Exposure to $PM_{2.5}$ levels above the city-specific 75th percentile reduces daily output quantity by 4%, which translates into a loss in output value by approximately $11 per developer compared to days with better air quality.

Chapter 2 provides causal evidence on the effect of in-utero exposure to air pollution on noncognitive ability in childhood. Noncognitive skills are important predictors for life outcomes like education, health and earnings. I exploit the meteorological phenomenon of thermal inversions to address the endogeneity in exposure to air pollution. To measure noncognitive abilities, I use data from a representative household survey in Germany on mother-reported Big Five personality traits assessed at ages 2-10. Pairing this with data on particulate matter (PM10) from outdoor monitors and reanalysis data on meteorological conditions, I find that an increase in particulate matter concentration by 1 $\mu g/m^3$ during the prenatal period raises neuroticism at age 5-10 by 7% of a standard deviation. This implies that affected children are less emotionally stable, more fearful and less self-confident. Back-of-the-envelope computations indicate that a one standard deviation increase in particulate matter reduces adult earnings by 0.24%-0.29% just through its impact on neuroticism.

Chapter 3, co-authored with Felix Holub and Ingo Isphording, studies how a misalignment between the circadian rhythm (the internal process that regulates the sleep–wake cycle) and social schedules affects the performance of high skilled knowledge workers. Using data from *GitHub* we measure hourly and daily output of roughly 40,000 software developers and classify them into morning- and evening-types based on their temporal activity profiles. We also observe with whom they collaborate on shared projects. We find that morning-types outperform evening-types, and this performance gap is driven by working days but not detectable on weekends or public holidays. Because prevailing social schedules are more aligned with natural rhythms of morning-types, this result suggests that a circadian misalignment reduces developer performance. Exploiting within-developer variation in composition of collaborators, we further show that output falls when developers have incentives to deviate from their natural rhythms in order to synchronize with collaborators in other time zones. Finally, we use a regression discontinuity design to investigate the impact of the transitions into and out of daylight saving time. We find that developer output drops after the spring transition, which introduces a discrepancy between social clocks and solar time. Together, these findings imply that a misalignment between circadian rhythm and social schedules impairs knowledge worker performance.

1. Air Quality, High-Skilled Worker Productivity and Adaptation

Evidence from GitHub

Joint with Felix Holub

1.1. Introduction

Driven by technological innovation, the world of work is undergoing rapid changes. Over the last decades, computerization has been causing an increase in the demand for workers performing non-routine, analytical, and interpersonal tasks (Autor et al., 2003; Autor and Price, 2013). Skills that complement digital technologies have been growing in importance: In the US, the share of jobs intensively requiring digital skills[1] has more than quadrupled from 5% to 23% between 2002 and 2016 (Muro et al., 2017). In parallel, there have been major shifts in the organization of work, complementing the growing role of IT. Teamwork, flexible schedules, and discretion in task choice are common, replacing fixed 9-to-5 schedules and direct task assignments, especially among highly-educated workers (Bresnahan et al., 2002; Mas and Pallais, 2020; Menon et al., 2020). Because jobs characterized by these task profiles, skill requirements, and organizational features form the backbone of the modern knowledge economy and are expected to become even more important as digitization and automation proceed, it is critical to understand what determines productivity in these settings.

In this paper, we study how environmental shocks impact performance and work patterns among highly skilled knowledge workers in a flexible work environment. Vast populations are exposed to environmental conditions such as heat and poor air quality, which have been shown to reduce labor productivity in several settings. Existing research, however, has considered routine jobs and/or inflexible work contexts (e.g. Chang et al., 2019; Somanathan et al., 2021). In the settings described above, workers not only use different skills, they also have flexibility

[1]Examples of digital skills are the abilities to handle information and communication technology and to conduct data analyses.

and discretion in organizing their workday, which might allow them to adapt to productivity shocks, thereby alleviating output effects. Moreover, in collaborative work settings, impacts of environmental shocks might get dampened, e.g., if co-workers can help each other to focus, or get amplified due to complementarities if co-workers depend on each others' input.

We focus on the effects of air pollution as it is a ubiquitous public health threat in urban areas across the globe.[2] 82% of the global population are exposed to levels of fine particulate matter exceeding World Health Organization (WHO) guidelines. To implement optimal air quality standards and policies to curb pollution, accurate estimates of the welfare costs of pollution are fundamental. Recent research shows that air pollution not only causes premature mortality and severe health damages, but also sub-clinical effects on labor market outcomes, student performance, or decision-making (see Aguilar-Gomez et al., 2022, for a review). The sub-clinical impacts play an important role for the total economic cost of air pollution as they affect a broad population, while morbidity and mortality effects are concentrated among vulnerable groups, like infants and the elderly.

We study the causal effect of air pollution exposure on professional software developers, using data from *GitHub* to measure developer output and work patterns. Software development is a STEM (science, technology, engineering, and math) occupation that requires analytical and advanced digital skills and generates high value for consumers, other industries, governments, and the research community.[3] Adverse productivity effects of air pollution in this occupation would thus have important implications for growth, innovation, and competitiveness. GitHub is the world's largest online code hosting platform, used for storing and jointly working on coding projects. It puts great emphasis on facilitating collaboration between developers. Moreover, software developers work in highly flexible settings that usually offer discretion over working hours and the tasks a developer chooses to work on at a given point in time. With these features, software development on GitHub is representative of the settings that characterize modern knowledge work.[4]

The GitHub data allow us to address the challenge that output of knowledge workers is often difficult to observe. We collect data on 25,000 users across four continents who work on projects owned by tech companies, indicating that they are professional software developers. The data includes users' locations as well as records of all actions they conduct in public projects along with precise timestamps and some further characteristics of the underlying task. We construct a user-by-day panel including measures of work quantity and quality,

[2]In an extension, we also provide some evidence on the effects of extreme temperatures for comparison.

[3]Median annual pay of software developers in the US was $110,140 in 2020 (Bureau of Labor Statistics, 2021).

[4]We provide evidence that software development is representative of modern high-skilled work in Appendix Figure 1.B.1 and Table 1.A.1: Required skills are similar (e.g., critical thinking and complex problem solving), except for substantially stronger digital skills like programming. Both software developers and high-skilled workers in general have a lot of flexibility in organizing their work and often work in teams.

working hours, and task choice for the period between January 2014 and May 2019. Based on developers' locations, we match these outcomes to city-level air quality monitor data on particulate matter smaller than 2.5 µm ($PM_{2.5}$). To account for endogeneity in air quality when estimating a model of developer activity, we follow Deryugina et al. (2019) by instrumenting $PM_{2.5}$ concentration with daily average wind direction. The 2SLS strategy exploits the effect of plausibly exogenous regional air pollution transport on local pollution levels, controlling for a wide range of other weather characteristics.

To measure daily output quantity, we count the total number of actions performed, including commits (individual code changes), opening and closing of pull requests and issues, and comments written in discussion fora.[5] As we can classify the different GitHub actions into *individual* and *interactive* activities (e.g., commits vs. comments), we can analyze heterogeneity in the effect of pollution exposure on performance in these two distinct types of work, which are both widespread in modern high-skilled jobs. To assess output quality, we compute the share of commits that get undone at a later point and the share of pull requests that get rejected as measures of error frequency. We also derive monetary estimates of the value of GitHub activities, exploiting additional data from an online marketplace where GitHub project owners offer payments for contributions to their projects. This allows us to translate the effects of pollution on output into monetary losses.

To study adaptation in flexible work arrangements, we investigate whether software developers exploit their discretion with respect to working hours and task choice in order to adjust to changes in air pollution. Specifically, we exploit information on the complexity of tasks addressed by developers' actions and on action timestamps to study whether they focus on easier tasks or adapt their working hours.

Our dataset covers 36 countries, including both developing and developed countries, with large variation in pollution levels, income, and pollution awareness across sample cities. We exploit this in heterogeneity analyses to explore how air pollution damages are distributed and to study the mechanisms underlying the pollution impacts.

We present three main findings. First, developers produce less output on days with higher levels of fine particulate matter. When $PM_{2.5}$ concentration reaches unusually high levels – exceeding the city-specific 75th percentile – the number of daily actions observed on GitHub falls by 4%. This effect is mainly driven by a decline in individual coding activity: The number of commits decreases by 6.2%. By contrast, collaborative or interactive work (e.g., commenting on issues) is much less affected. Compared to other occupations studied in previous research, including both physically- and cognitively-demanding jobs, the magnitude of the effects on output is relatively small. Nonetheless, the pollution-induced output declines translate into

[5]Pull requests are a tool to suggest changes to the code base of a repository, for more details see Section 1.3.

relevant monetary damages due to the high value generated by software developers. The loss in output value per developer and day amounts to \$4 for a standard deviation increase in $PM_{2.5}$, i.e., common fluctuations in pollution, and \$11 for days with unusually high air pollution.

Second, output quality is unaffected by changes in air pollution. We find no evidence that software developers commit more errors on days with high levels of $PM_{2.5}$. A potential explanation for the absence of quality effects, as well as the modest size of effects on quantity, is worker adaptation to pollution-induced productivity shocks in this flexible work setting.

Our third main result provides evidence for this: We find that developers exploit their flexibility in task choice by focusing on less complex tasks when air pollution increases. Among activities related to issues[6], the share that refers to issues labeled as relatively easy increases by 5% on days with $PM_{2.5}$ concentration above the city-specific 75th percentile compared to low pollution days. Similarly, code submitted or reviewed by developers changes 4% fewer files and contains 2% fewer new lines, indicating that the code addresses less complex tasks. Among developers with stronger adjustment in task choice in response to $PM_{2.5}$ exposure, effects on output quantity are attenuated.

Software developers also adapt their working hours. They reallocate work activity from high-pollution, low-productivity days to low-pollution, high-productivity weekends. In particular, developers end work activity earlier on days with unusually high $PM_{2.5}$ concentration. To compensate, they work more on weekends after a workweek with poor air quality, especially if pollution concentration on the weekend is moderate, i.e., below the city-specific 75th percentile. The increase in weekend work makes up for 33% of the reduction in coding output on the day of exposure. These forms of adaptation likely explain why, compared to other professions, we find moderate effects of particulate matter on output. At the same time, compensation by working more on the weekend also implies an additional welfare cost due to forgone leisure time and potential negative impacts on the work-life balance. Losses in output value are thus likely a lower bound on the overall cost of air pollution in this setting.

The adverse effects of pollution on output quantity arise at concentrations below the regulatory standards in force in the European Union and the US. Indeed, effects are strongest at low levels of $PM_{2.5}$. The effect magnitude does not vary systematically with country-level pollution awareness, indicating that the negative effects on output are not driven by avoidance behavior. This is corroborated by the fact that we find relatively small extensive margin effects. We also show that effects are substantially larger in locations with an older building stock, suggesting that differences in effective indoor pollution exposure play an important role. This points towards a physiological mechanism underlying our main results.

In a dynamic analysis, we show that pollution exposure on a given day reduces both con-

[6] This includes creating, closing, and commenting on an issue.

temporaneous output and, to a lesser extent, output produced over the following two to three days, but not thereafter. We repeat our main analysis at the monthly level, to assess the effect of pollution on output net of adjustment and accounting for dynamic impacts. We again find negative effects on the number of actions performed, driven by less individual coding activity. The impact of an increase in daily $PM_{2.5}$ concentration by one unit is roughly 60% larger according to these results compared to the results at the daily level, but still moderate in comparison to other professions. A standard deviation increase in daily $PM_{2.5}$ generates a loss in monthly output value of approximately \$6. In the monthly analysis, we also consider the growth rate of a developer's followers on GitHub as a summary measure of work quantity, quality, and relevance. Air pollution exposure also reduces this outcome, indicating that it not only reduces short-run performance but also slows down the build-up of reputation, which could plausibly have long-run consequences for career paths.

Overall, our results imply that improvements in air quality generate economic benefits in terms of productivity gains among highly skilled STEM workers. While flexible work arrangements with respect to schedules and task choice allow workers to adapt to productivity shocks, even in the very flexible setting we analyze here, air pollution generates economically relevant costs. This is true even for locations with relatively low pollution levels in compliance with existing regulatory standards.

Related Literature. This paper contributes to the research on the effect of environmental factors on economic outcomes. Firstly, our paper directly links to the literature strand on air pollution and worker productivity. Several studies document a negative impact of pollution on productivity in manual and routine occupations, such as textile workers (Adhvaryu et al., 2022; He et al., 2019), pear packers (Chang et al., 2016), call center agents (Chang et al., 2019), or fruit pickers (Graff Zivin and Neidell, 2012). A small number of papers also demonstrate negative effects of poor air quality on performance in more cognitively-demanding occupations. This evidence comes from studies on error detection of baseball umpires in the US (Archsmith et al., 2018), the speech quality of Canadian politicians (Heyes et al., 2019), and case handling time by Chinese and Mexican judges (Kahn and Li, 2020; Sarmiento, 2022).

While these contexts allow to create precise measures of worker performance in a specific domain, the settings analyzed do not reflect the typical features of work organization in most modern high-skill jobs. As outlined above, frequent collaboration, multi-tasking, and flexibility in work schedules are widespread in these jobs and each of these characteristics might affect the severity of pollution-induced productivity shocks. Furthermore, the rather inflexible settings studied so far do not allow to analyze worker adaptation to pollution. Related work investigates performance in cognitively-demanding tasks, but outside of standard work settings, e.g., among chess players (Künn et al., forthcoming), individual investors (Huang et al., 2020), or brain game players (La Nauze and Severnini, 2021; Krebs and Luechinger, 2021).

We contribute to this literature by expanding the analysis to a STEM profession that is representative for a large group of high-skilled workers in flexible, modern work environments. Thus, our analysis adds novel insights into the labor market cost of air pollution that will likely still be relevant after future waves of digitalization. We present first evidence on productivity effects separately for individual and collaborative activities, a distinction absent from previous work. In addition, while existing papers are based on data from a single country and often only a single site, we work with a large sample of developers across multiple countries. This allows to draw more general conclusions about the pollution-productivity relationship and to explore effect heterogeneity, e.g., with respect to local income levels or pollution awareness.[7]

Secondly, we contribute to the literature on worker adaptation to environmental shocks and connect it to research on the effect of flexible work arrangements on productivity. A number of papers study how workers adjust working hours in response to temperature shocks (Graff Zivin and Neidell, 2014; Neidell et al., 2021; LoPalo, forthcoming). With respect to air pollution shocks, Adhvaryu et al. (2022) and Bassi et al. (2021) demonstrate an important role of managers who can mitigate productivity losses, e.g., by reallocating workers to different tasks. These studies, however, focus on rather low-skilled manufacturing workers. Our work identifies a new margin of adjustment in a flexible high-skilled setting, namely task choice. Moreover, we present new evidence on temporal reallocation of work activity towards the weekend in response to pollution shocks.[8] By showing that workers exploit discretion in task choice and working hours to adapt to an environmental shock, and thereby alleviate its impact on productivity, our work links to research on the causal effects of flexible work arrangements and worker autonomy on productivity. Shepard et al. (1996), Beckmann et al. (2017), and Angelici and Profeta (2020), for instance, find across different contexts in the US, Germany, and Italy that working time autonomy increases employee productivity. Our results suggest that the ability to adapt to idiosyncratic productivity shocks might contribute to the positive relationship between flexible work arrangements and performance.

Our analysis of worker output and behavior also relates to a broader literature that studies drivers of worker productivity. For example, Lazear et al. (2015) show that productivity increased during recessions because of higher worker effort. Pencavel (2015) and Shangguan

[7]Borgschulte et al. (forthcoming) and Fu et al. (2021) do not consider specific professions, but conduct broader analyses of air pollution and labor earnings in the US and manufacturing sector productivity in China, respectively. We add new evidence relative to these papers due to our international sample and our analysis of worker adaptation which requires high-frequency microdata.

[8]In parallel work, Hoffmann and Rud (2022) also show evidence that workers reallocate labor supply across days in response to changes in $PM_{2.5}$. However, they study a different setting, namely formal and informal workers in Mexico City, whereas we focus on a sample of highly-skilled STEM workers. Moreover, Hoffmann and Rud (2022) interpret the temporal substitution as a strategy to avoid pollution exposure, whereas in our setting it serves as compensation for reduced productivity.

et al. (2021) examine how the output of workers is driven by their work hours. Kaur et al. (2021) find that financial concerns reduce productivity via psychological channels. We complement this literature with detailed evidence on how environmental shocks affect work patterns and performance.

Lastly, another contribution of this paper is to demonstrate new ways to use publicly available data on GitHub activity. While we are not the first to use this data in economics,[9] we propose strategies to construct a sample of highly active users who are likely professional software developers, to study task difficulty, and to estimate the monetary value of the output observed on GitHub.

Outline. The remainder of the chapter is organized as follows: We begin in Section 1.2 with a short explanation on how pollution affects the human body. Section 1.3 follows with a description of *Github*, the data and sample. We explain the research design and how we implement the two-stages least squares strategy of Deryugina et al. (2019) in Section 1.4. Our main results are presented in Section 1.5. Findings from heterogeneity analysis and extensions follow in Section 1.6. Section 1.7 concludes.

1.2. Background on Particulate Matter

In our analysis, we focus on $PM_{2.5}$ to measure air pollution, i.e., particulate matter with a diameter of less than 2.5 µm. Particulate matter refers to all solid and liquid particles suspended in the air. In most urban areas around the world, the majority of $PM_{2.5}$ originates from anthropogenic sources, including traffic, industrial production and biomass burning (Karagulian et al., 2015). While some $PM_{2.5}$ is produced locally, for example by traffic, in most areas a significant share of local pollution arises from distant sources via long-range transport. Power plants and industrial facilities generate precursor emissions, e.g., sulphur dioxide (SO_2), which are transformed into secondary particulate matter, e.g., sulfate (SO_4^{2-}) and can get transported over long distances (Almeida et al., 2020).

A key reason to focus on $PM_{2.5}$ in our study is the fact that these small particles can penetrate indoors and are thus of major relevance for indoor office workers. Deng et al. (2017) find indoor-outdoor ratios of $PM_{2.5}$ between 0.4 and 1.2 for office and apartment buildings in Beijing, which can only be reduced to a level near zero with high-quality indoor air cleaning systems. In line with that, Xu et al. (2020) and Hoek et al. (2008) report significant and sizeable correlations between indoor and outdoor fine particulate matter for other cities in China and Europe.

[9] McDermott and Hansen (2021), e.g., use the data for an analysis of the impacts of the COVID-19 pandemic on work patterns.

Moreover, a large body of research documents that fine particulate matter plays a key role for the adverse effects that air pollution exerts on various dimensions of human health. The small particles can penetrate deeply into the lungs, causing damage to the respiratory system, including reduced lung function, asthma, and chronic obstructive pulmonary disease. Similarly, epidemiological and economic studies find evidence for cardiovascular health effects like high blood pressure and heart diseases (Lederer et al., 2021). Medical research on humans and animals points to systemic oxidative stress, inflammation and endothelial dysfunction (impaired functioning of the inner lining of blood vessels) as underlying biological mechanisms (Anderson et al., 2011; Kelly and Fussell, 2015). While severe morbidity and mortality effects are concentrated among vulnerable groups like the elderly and infants, even healthy individuals can experience mild, subclinical effects, including irritation in the nose and throat or coughing (Pope, 2000). In response to the mounting evidence on adverse health effects, several countries have introduced standards on annual ambient PM2.5 concentrations, and typically tightened them over time. Currently, standards are in force e.g. in the US (12 $\mu g/m^3$), Canada (10 $\mu g/m^3$) and the European Union (25 $\mu g/m^3$). The WHO guidelines recommend a level of no more than 5 $\mu g/m^3$.

Recent clinical and epidemiological studies imply that exposure to fine particles also exerts adverse effects on the central nervous system (Delgado-Saborit et al., 2021; Babadjouni et al., 2017). Small particles have been found to reach the brain via the olfactory pathways and the bloodstream. Animal and autopsy studies indicate that particulate matter causes neuro-inflammation, which can lead to cognitive impairments and neuro-degenerative processes (Calderón-Garcidueñas et al., 2007). Associations between pollution exposure and changes in brain structure have been detected in neuroimaging studies. Consistent with this, La Nauze and Severnini (2021) find that brain game players score 0.18 standard deviations lower when $PM_{2.5}$ concentration exceeds 25 $\mu g/m^3$ as compared to days with better air quality. Similarly, Ebenstein et al. (2016) find that short-run exposure to $PM_{2.5}$ reduces student performance on high-stake exams.

Overall, the research on particulate matter, health, and cognitive functioning implies that $PM_{2.5}$ exposure might plausibly reduce productivity both in physically- as well as cognitively-demanding tasks. Growing evidence in economics on negative productivity effects in manual occupations, and on reduced cognitive performance confirms this. We intend to *quantify* productivity impacts in a high-skilled work environment, and investigate potential adaptation responses that might occur in these settings.

1.3. Setting and Data

To analyze the effects of air pollution on productivity and work patterns in a high-skilled profession, we pair information on *GitHub* activity for a global sample of software developers with data on local air quality. This is complemented with data on meteorological conditions to construct the instrumental variables and to control for local weather. This section starts with a brief description of GitHub, followed by an overview of the GitHub data and how we use it to measure developers' productivity. After checking the validity of these outcome measures, we end with a description of the environmental data.

1.3.1. Setting: GitHub

GitHub is built on *Git*, an open source version control system that records who changed which part of a file at what point in time. *GitHub* is a web platform for hosting Git repositories[10] and, on top of the version control functionality, provides additional features to facilitate collaboration. For each of their repositories (or repos for short), owners can choose whether to make it public or private, i.e., whether the respective files are visible to everyone, or only to the repository members. In 2019, more than 30 million accounts were registered on GitHub, who together owned more than 120 million public repositories, making it the world's largest host of source code.

The core action in Git is a *commit*, which refers to saving the current version of the repository after implementing a change to a file, or a set of files. As such, a commit represents that some work on code files was conducted by the commit author. Only repository owners and team members invited by them can modify files via commits.

The primary additional collaboration features offered on GitHub are *pull requests* and *issues*. A pull request (PR) is a tool to propose code changes to a repository. To create a PR, a user generates a copy of the repository in question, implements the changes in his copy via commits, and then submits these to the original repo.[11] Repo members then review the suggested changes and decide whether to accept (i.e., merge) or reject them. Each PR includes a discussion forum where users can comment directly on the proposed changes. Feedback provided there can be implemented within the same PR. Due to these features, PRs facilitate collaborative coding and are thus not only used to contribute to projects of which a user is no member but also within project teams.

Issues are text messages, typically used to suggest improvements and organize tasks in a given repo.[12] Like PRs, issues contain a discussion forum where users can comment on the

[10]The term repository refers to the location where all files belonging to a project are stored.
[11]An example for a PR can be found at https://github.com/microsoft/vscode/pull/54244.
[12]For an example of an issue, see for instance https://github.com/microsoft/vscode/issues/39526.

problem or question at hand. Repository members can assign labels to issues in order to highlight their category (e.g., bug, feature request), priority, or difficulty. The platform provides nine default labels, and repository teams can create additional ones specific to their repo. Once an issue is resolved, it can be marked as closed.

On top of that, GitHub provides social network functions, e.g., options to follow other users and subscribe to specific repositories and issues to receive notifications about new activities.

1.3.2. GitHub Data on Productivity and Work Patterns

GitHub actions related to commits, PRs, and issues reflect productive work aimed at building or improving software products. Hence, we collect data on these activities to measure output generated by highly skilled developers. The GHTorrent project provides information on GitHub users and all actions they conduct in *public* repositories in the form of a relational SQL database. We use the version of the database containing data up to June 1st, 2019. The user table comprises a unique identifier, login name and registration date for all users registered on the platform at this point. In addition, location and company information as stated on the user profile on this date is reported. The projects table provides identifiers and names of all public repositories as well as a reference to the user owning the repo. Data on activities is available separately by type of action (e.g., commits, opening issues, PR comments, etc.) and includes the exact timestamps and the identifiers of the acting user and the repository where the event was conducted in. For specific actions, further information is reported, e.g., the labels attached to issues.

We complement this with data from GHArchive, which also provides a record of actions in public repositories, and contains additional information on some events, e.g., the title of a commit (called commit message), and the number of lines of code added and deleted, as well as the number of files changed within a PR. GHArchive and GHTorrent data can be linked via users' login names.

These data have multiple favorable features for our analysis. First, the precise records of activities conducted on GitHub enable us to quantify daily output produced by software developers. In this way, we address the long-standing challenge that work conducted by high-skilled workers during a given period is often difficult to measure. Second, the data cover all GitHub users which gives us a much broader geographic coverage and thus a clear advantage in terms of external validity compared to previous studies based on data from only one country, and often even just one sampling site. Moreover, the rich information included allows us to investigate not only changes in output quantity, but also quality and work patterns, which are of major relevance in high-skill professions.

The data also have clear limitations. In order to assign local air quality to users, we rely on self-reported locations. Some users might report wrong or outdated locations, giving rise to measurement error. Under the assumption of classical measurement error which is not correlated with pollution levels, this issue leads to attenuation bias such that any adverse effects we find can be considered a lower bound on the true effect. Additionally, we have no information on work conducted in private repositories or outside the platform. Many GitHub users conduct no or only little work in public repositories such that it would be impossible to identify any productivity effects of air pollution exposure based on their activity data. Thus, when constructing our analysis sample, we aim at capturing users who are professional software developers and do a substantial part of their formal work in public GitHub repositories.

Sample Construction. We focus on non-organizational users who report a location at the city level, which is the degree of geographic precision required to assign local air quality. Among them, we keep only users who ever committed in a repository owned by a company, i.e., users with the authority to change the source code of a company-owned project. This step is intended to focus on professionals who are in some way affiliated with the companies. To identify these users, we compile a list of repositories operated by companies[13] and then use the information on the repository where a commit was made from the GHTorrent data. To drop bots, we discard a small number of users with extremely high activity levels, very regular commit patterns, or suspicious login names.[14] To focus on cases where we can observe a substantial part of an individual's total work, we only admit users into the sample once they have at least 20 commits in public repos in a given month. They enter the sample in the month after they have passed this threshold for the first time. Users stay in the sample until the end of the observation period unless they conduct less than three *unproductive* actions in a given month. In this case, we drop users from the sample for that month, assuming that they might have moved to a different platform or work on projects in private repositories. Unproductive actions are activities we do not use as outcome variables, based primarily on the social network functions GitHub offers.[15] Lastly, we restrict the sample to users living in cities with at least 15 relevant users that are covered by our data on air pollution. This yields a sample of 27,701 users in 220 cities across 47 countries during the sample period from February 2014 until May

[13]This is based on a publicly available list of firms active on GitHub, which can be accessed at https://github.com/d2s/companies/blob/master/src/index.md and on the lists of open-source projects operated by Google, Microsoft and Facebook published on their web pages.

[14]Bots are computer programs typically used to automate specific tasks. On GitHub, some company-affiliated projects for instance employ bots to comment on newly opened issues and PRs to ask users to provide specific information on their issues or to sign a contributor license agreement. To make sure not to capture bots in our sample, we drop users if the number of actions or commits conducted by them is in the top 0.1 percentile of the distribution, if more than 20% of their commits occur at full hours (indicating automated commits) or if their login name indicates that they are bots.

[15]The unproductive actions include following another user, watching a repository, (un)subscribing to an issue, labeling an issue, and (un)assigning an issue to a user.

2019.[16] These locations are depicted in Figure 1.1. For the IV approach based on changes in wind direction, we require multiple cities in close geographic proximity, as outlined further in Section 1.4. Hence, in our main analysis, we focus on 193 cities across 36 countries depicted in dark blue, comprising 24,534 users. All descriptives reported in this section refer to this main analysis sample. In extensions, we also include the cities depicted in light blue.

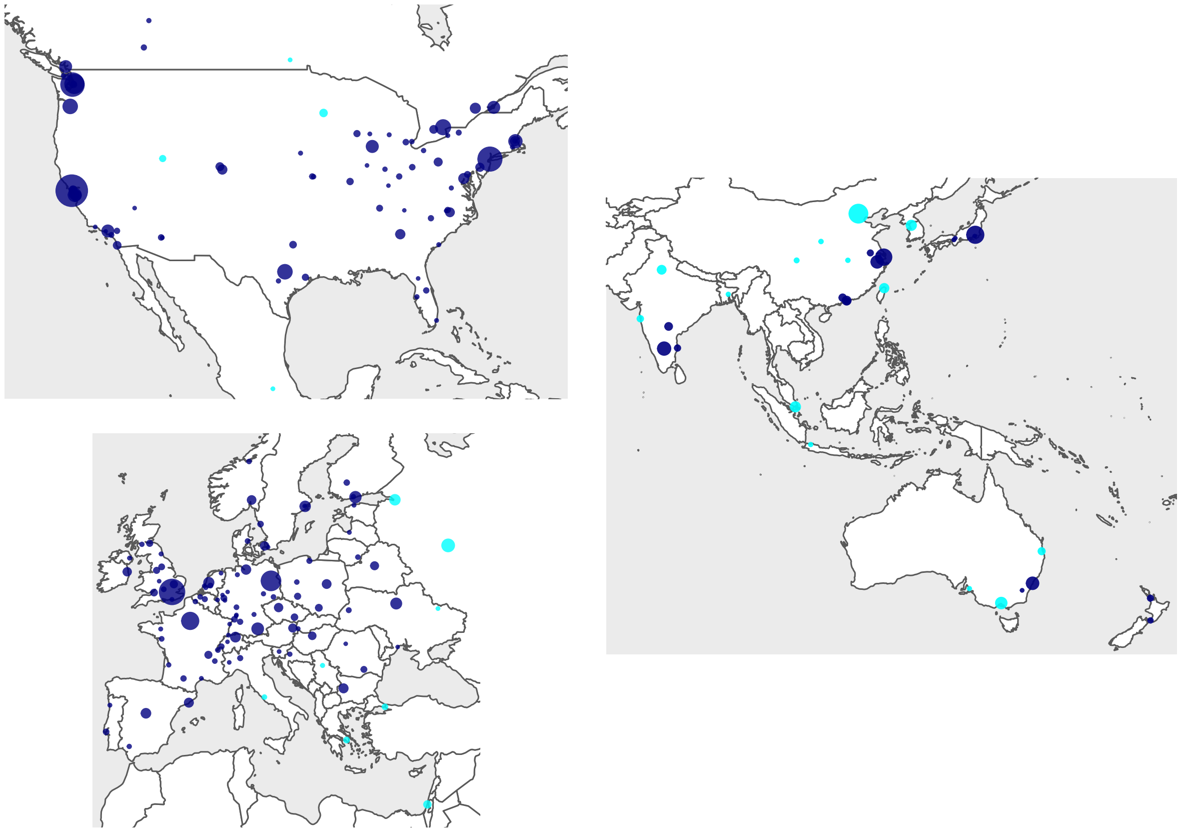


Figure 1.1.: Sample Cities

Note: Each point represents one sample city. Circle size is based on the number of users observed in the city. Cities depicted in dark blue are used in the main analysis and cities depicted in light blue are added in extensions.

Outcome Measures. For the analysis sample, we compile an unbalanced user-by-day panel including measures of output quantity and quality, as well as work patterns. To measure overall daily work *quantity*, we count the total number of productive actions conducted per user and day, after translating timestamps from UTC into local time. This is the sum of commits,

[16]During our sample period some users changed their location. Since the GHTorrent data on users is a snapshot taken on June 1st, 2019, we use earlier versions of the database (one snapshot in each year between 2015 and 2018) to check for movements. In total, 6.3% of users reported more than one distinct location during this period. We identify the city where they spend the biggest part of the sample period, and keep them in the sample only while they resided in this city.

comments written on PRs, issues or commits, creations of PRs and issues, closing of PRs and issues, and reopening of issues. Furthermore, we separately count the number of commits and comments because these two action categories are observed most frequently on GitHub and reflect two distinct types of work. While the number of commits provides a measure of individual coding activity, comments reflect participation in discussions about issues and code changes, i.e., collaborative work. This allows us to conduct the first analysis of the productivity impacts of air pollution in a high-skill profession that takes potential effect heterogeneity between individual and interactive work into account.[17]

To assess the *quality* of users' output, we measure the share of all PRs opened on a given day that are merged, i.e., accepted. PR rejection points to issues in the code or style. Secondly, we compute the share of commits made by a user on a given day that were reverted at a later point. Reverting a commit, i.e. removing all changes made in it, indicates severe problems that cannot easily be fixed in follow-on commits.[18]

To analyze worker *adaptation*, we build measures of task choice and working hours. Firstly, to explore whether users switch to easier tasks on high-pollution days, we leverage information on the complexity of issues and PRs. In the case of PRs, we use the number of new lines of code added, of lines of code deleted, and of code files changed as measures of their complexity. We take the average value of these variables across all PRs a user worked on a given day, either by creating, reviewing, or commenting on the respective PR. To assess the difficulty of issues, we rely on the user-assigned issue labels. We exploit the fact that there are several labels indicating that a given issue is relatively easy, e.g., the default labels *good first issue*[19] and *documentation*,[20] or individual labels such as *beginner friendly* or *low-hanging fruit*. The complete list of labels we use to identify easy tasks is depicted in Appendix Table 1.A.2. We construct the share of all issue events conducted by the user (commenting, opening, closing, or reopening of issues) which refer to an *easy* issue. With this approach, we do not have to evaluate issue complexity ourselves but can rely on the assessment by experts who know the project in question very well. Furthermore, the label is visible to all users, i.e., workers

[17] The other action types occur less frequently and are thus considered as secondary outcomes. The number of PRs created also reflects individual work on code, whereas the number of issues closed, opened, or reopened provides additional measures of interactive work. The number of PRs closed reflects code review and decision-making on whether to merge or reject the proposed changes.

[18] The act of reverting a commit is itself a commit, which has a specific, auto-generated commit message. The messages are available in the GHArchive data and allow to identify both revert commits and the original commit that is reverted.

[19] This issue was introduced by GitHub to encourage first-time contributions, but does not imply that the issue cannot be addressed by more experienced developers.

[20] The documentation label is included because work on the documentation is typically easier than work on code to fix bugs or build new features. This follows, e.g., from Tan et al. (2020) and from the fact that GitHub also used the documentation label in their approach to construct the good first issue label (for details see https://github.blog/2020-01-22-how-we-built-good-first-issues/).

searching for easy tasks due to an adverse productivity shock can easily identify these issues as suitable.

Secondly, we exploit the timestamps reported in the data to approximate users' working hours in order to investigate whether they try to make up for their reduced productivity by working longer hours in the evening or on the weekend. Evening work is measured by the minute of the day the last action of the day was performed and the share of actions conducted after standard working hours, i.e., after 6 pm. To measure weekend work, we use the sum of actions conducted on Saturday and Sunday of each week.

Finally, as a summary measure of the quantity, quality, and relevance of a user's work, we consider the monthly growth rate of the number of a user's followers. This allows us to investigate whether air pollution exerts only temporary effects on daily output, or whether it also generates effects on users' reputation and influence over a longer time horizon.[21]

Descriptives. Table 1.1 presents summary statistics on the outcome variables. On average, users perform 2.77 actions per day, of which 1.29 are commits and 0.93 are comments. The remaining productive GitHub actions—opening and closing issues and PRs—are observed less often. Hence, the sample users are indeed highly active in public repos, as casual users who work on GitHub only occasionally can hardly achieve such figures, especially given that we average across all days, including weekends and holidays.[22] On average, users are active on 37% of all days in the sample period. Conditional on being active at all, the mean number of actions per day is 7.59. A commit reversal is a very rare event, indicating severe errors. It happens among only 0.2% of all commits made in the sample. Rejection of a PR occurs more frequently, in 33% of all cases. On average, 7% of all issue events refer to an easy issue. 29% of actions are made after 6 pm. The mean time of the final action of the day is 5:45 pm.[23]

Figure 1.2 provides more detailed information on the distribution of activity across days of the week and hours of the day. The solid lines depict the share of all activity that is conducted during the respective hour of the day on weekdays (left) or weekends (right), respectively. We present this share for commits, comments, and total actions. Activity levels are highest during core working hours (marked in grey) and considerably lower in the evening and night hours and on weekends. Notable activity during evening hours and on weekends is not uncommon among highly educated workers (Mas and Pallais, 2020). Overall, the distribution is similar across all three variables. However, comments, i.e., interactive activities, are even more

[21] We provide a list of all outcomes and details on their construction in Appendix Table 1.A.3.

[22] Overall, our main sample comprises only 0.076% of all GitHub users but accounts for a disproportionally large share of all activities in public repositories, e.g., 2.1% of issue creations, 7.6% of issue closings, 9.9% of comments written and 6.6% of PR actions (opening and closing).

[23] To take into account that high-skill workers often work long hours in the evening, we define a work day to last from 3 am on the calendar date to 3 am on the following day.

concentrated during standard working hours as compared to commits, i.e., individual coding activities. This is plausible given that the more collaborative tasks are more productive during common working hours, when other users are working as well.

Table 1.1.: Summary Statistics for the Analysis Sample of GitHub Users

	Mean	SD	SD (within)	Min	Max	Observations
Output Quantity						
Actions	2.77	7.26	6.46	0	293	14,538,351
of which Commits	1.29	3.82	3.55	0	234	14,538,351
Comments	0.93	3.37	2.94	0	280	14,538,351
PRs opened	0.15	0.71	0.67	0	151	14,538,351
Issues opened	0.10	0.80	0.74	0	222	14,538,351
PRs closed	0.17	0.94	0.89	0	284	14,538,351
Issues closed	0.12	0.89	0.87	0	263	14,538,351
Any action	0.37	0.48	0.44	0	1	14,538,351
Actions | Actions > 0	7.59	10.39	9.22	1	293	5,310,794
Output Quality						
Share PRs merged	0.67	0.45	0.40	0	1.0	1,132,824
Share commits reverted	0.002	0.029	0.029	0	1	3,458,932
Task Complexity						
Share easy issue events	0.07	0.21	0.20	0	1.0	3,203,772
Files changed per PR	9.37	60.15	60.00	0	9660	1,738,027
Lines added per PR	347.05	1660.15	1618.98	0	64425	1,738,027
Lines deleted per PR	125.56	700.38	686.17	0	25037	1,738,027
Working Hours						
Share actions after 6 pm	0.29	0.39	0.36	0	1.0	5,296,035
Time last action	17:45	5.01 hours	4.63 hours	3:00	3:00	5,296,035

Note: This table describes the main analysis sample at the developer×date level. The first two panels provide summary statistics for outcome variables we use to measure output quantity and quality. The bottom panels describe variables measuring task complexity and working hours. The table displays the mean, standard deviation, within-developer standard deviation, minimum and maximum value of the variables as well as the number of observations.

Finally, Figure 1.3 presents information about the work status of users in our sample. The left plot depicts the most frequent terms used in the biographies (bios) on their GitHub profiles. 36% (9,507 users) of the sample provide such a self-description. The data was accessed via the GitHub API. For each term, we measure in what share of all bios it occurs, after stemming and removing stop words. Three terms clearly stand out: engineer/engineering, software, and developer/development occur in 15% to 25% of all bios, much more often than any other words. The right plot complements this with information on employers which users can report on their GitHub profiles. In our sample, 61% (16,385 users) provide some information in this field

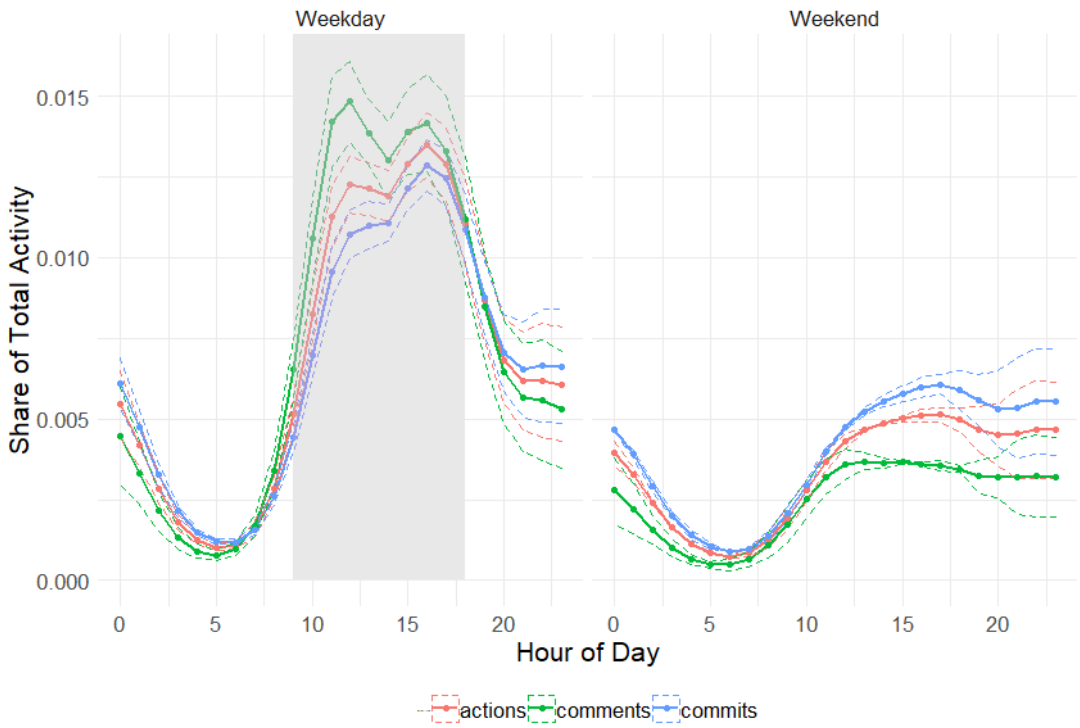


Figure 1.2.: Distribution of Activity across Hours of the Day and Days of the Week

Note: Solid lines reflect the share of activities by sample users which is conducted during the respective hour of the day on weekdays (left) or weekends (right), separately for total actions, commits and comments. Dashed lines mark 95%-confidence intervals. Grey area reflects core working hours, 9 am to 6 pm on weekdays.

with Microsoft and Google being the most frequent employers, followed by Facebook and Red Hat, i.e., big US-based tech companies strongly engaged in open-source. While we are unable to assess whether the subsample of users who provide a bio or the company information is representative, the clear peaks in the two plots at work-related terms and well-known tech companies, together with the concentration of activity in core working hours, strongly suggest that we do capture professional software developers who use GitHub as part of their formal work. Thus, in the remainder of the paper, we use the term ‘developers’ when referring to the users in our sample.

1.3.3. Gitcoin: Monetary Value of GitHub Activity

To assess the validity of the productivity metrics constructed from the GitHub data and to translate the estimated effects of air pollution on these outcomes into monetary damages, we draw on data from a platform called *Gitcoin.*[24] Two types of agents interact on this platform:

[24]Gitcoin was founded in 2019 and is complementary to GitHub. At the end of 2021, about 300,000 GitHub users were registered on the platform.

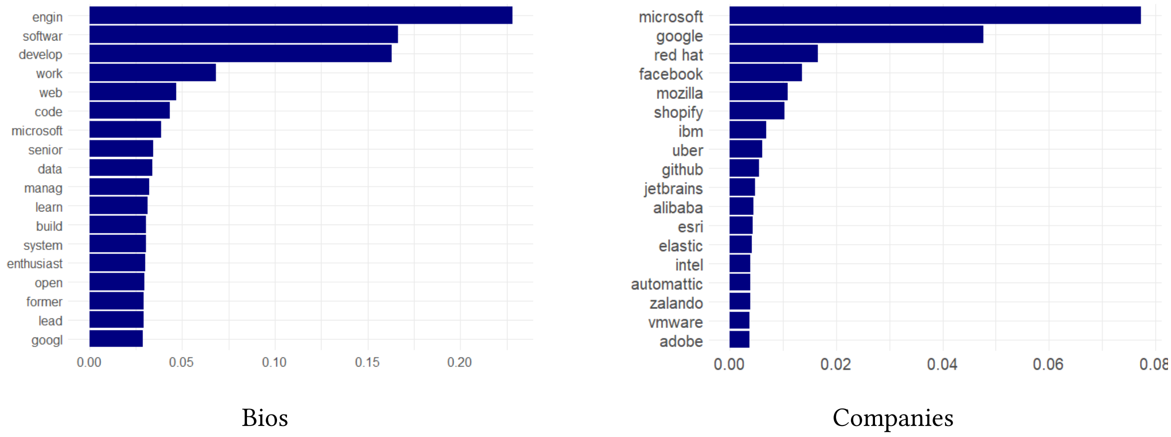


Figure 1.3.: Most Frequent Terms from User Self-Descriptions and Company Fields

Note: The left bar plot is based on data from 9,507 user bios, accessed in 2021 via the GitHub API. Words in the bios are transferred to lowercase, stemmed, and stop words are removed. The total word count is divided by the number of bios. The plot on the right is based on data from June 2019 on 16,385 users, collected from the company column in the GHTorrent user table.

GitHub project teams aiming to incentivize external contributions to their projects post open issues from their public repos on Gitcoin, along with information on issue characteristics and a payment they offer for a solution. On the other side of the transaction, freelance developers can apply to solve these issues and earn money for their contributions.

Work on the issues is submitted in the form of a PR in the respective GitHub repo. If the PR is accepted by the issue funder, the PR author receives the payment, typically in cryptocurrencies. We collect data on 292 issues for which PRs were submitted and payments made by March 2022 via the Gitcoin API, including the value of the payment in USD and the hours worked on the PR as reported by the submitting user. We merge this with information on the size of the respective pull request obtained via the GitHub API (number of commits, number of lines of code added and deleted, and number of files changed). A detailed description of the data is provided in Appendix 1.C.

In the data, we find mean payments of $354 per pull request and $112 per commit, one of our primary outcome variables. On average, developers spend 1.8 hours on one commit. Hence, their implied hourly wage amounts to $62, almost coinciding with the mean hourly wage of $58 among software developers in the US in 2021 as reported by the Bureau of Labor Statistics (2021). We will use the monetary values of commits and PRs, our measures of individual coding activity, to translate the estimated effects of air pollution on these outcomes into monetary damages.

Are the outcomes we consider valid measures of productivity and task complexity? In Appendix Tables 1.C.1 to 1.C.3 we use the Gitcoin data to test this. We find that both the payment awarded for the PR and the time spent on creating it are consistently positively correlated with the number of commits the PR comprises. We view this as a confirmation that changes in the number of commits reflect fluctuations in developer productivity. Holding the number of commits constant, adding more lines of code and changing more files in a PR is associated with a higher payment, suggesting that these variables indeed reflect task complexity. 12% of the Gitcoin issues are labeled as *easy* according to our definition. PRs addressing these issues are on average rewarded $186, only half the amount among PRs addressing other issues. Even holding all aforementioned PR characteristics constant, these issue labels are negatively associated with the value of a PR, supporting their validity as indicators for easy tasks.

1.3.4. Environmental Data

Air Quality. The pollutant of interest in our analysis is $PM_{2.5}$. We collected data on $PM_{2.5}$ concentration measured at outdoor monitors in and around the sample cities from several environmental agencies. For a few cities, we could not obtain monitor-measured data but instead used high-resolution reanalysis data from the Copernicus Atmosphere Monitoring Service (CAMS). Reanalysis datasets are constructed by combining measurements taken at ground-level monitors, satellite images, and atmospheric transport models. Appendix Table 1.A.4 provides a detailed list of the data sources. All data is provided at either the daily or the hourly level. Where necessary, we transfer hourly data into local time and aggregate to the daily level. Cities are assigned the simple average of all available monitor readings within a 40km radius around the city centroid.[25] Our data on $PM_{2.5}$ covers 96% of all city×day observations.

We winsorize $PM_{2.5}$ at the continent-specific 0.1^{th} percentile and the 99.9^{th} percentile to ensure that our results are not driven by extreme outliers (e.g., extremely high concentration of fine particulate matter due to heavy wildfire smoke). The population-weighted average $PM_{2.5}$ concentration in the sample is 12.4 μg/m^3 (standard deviation: 14.5 μg/m^3, within-city: 11.8 μg/m^3), i.e., slightly above the annual standard of 12 μg/m^3 set by the U.S. Environmental Protection Agency (EPA) and clearly above the WHO guideline value for the annual mean concentration (5 μg/m^3). Figure 1.4 displays the distribution of daily $PM_{2.5}$ concentrations in our sample, separated by seven large geographic regions, $R \in$ {Northern Europe, Southern Europe, Western Europe, Eastern Europe, North America, Oceania, Asia}.[26] Air quality exhibits substantial heterogeneity across regions: Cities in North America, Oceania, and Northern Europe have relatively clean air, with concentrations above 20 μg/m^3 rarely being observed. Locations in Southern and Eastern Europe by contrast experience this level of pollution on 28% of

[25] CAMS reanalysis data is reported on a 0.1° longitude×0.1° latitude grid. Given the large number of grid points, we only use measurement points within 25km of the centroids for the relevant cities.

[26] We show the distribution of observations in our developer×date panel across these regions in Table 1.A.5

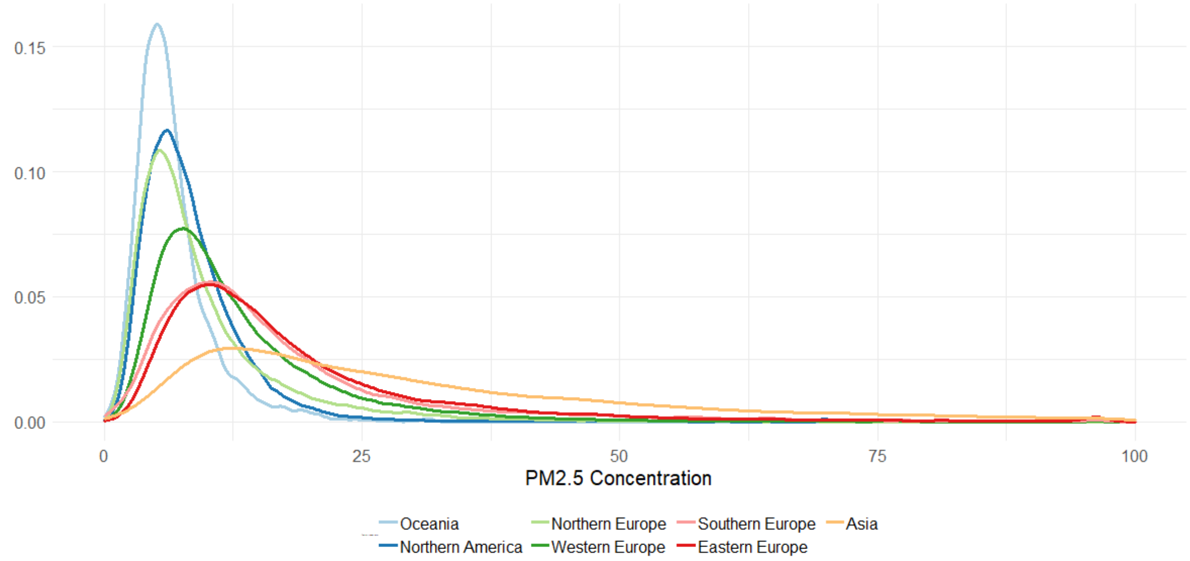


Figure 1.4.: Distribution of Daily $PM_{2.5}$ Concentrations by Geographic Region

Note: The plot shows densities of $PM_{2.5}$ concentration based on 353,445 city×day observations, separately by geographic region. Oceania: Australia, New Zealand. Northern Europe: Scandinavia, UK, Ireland, and the Baltic countries. Southern Europe: Portugal, Spain, Italy, Croatia, Slovenia. Asia: China, India, Japan, Hong Kong. Northern America: US, Canada. Western Europe: Switzerland, Austria, France, Germany, Belgium, and the Netherlands. Eastern Europe: Poland, Czech Republic, Hungary, Belarus, Ukraine, Slovakia, Bulgaria, Romania.

all days, and Asian cities even 60% of the time.

Wind conditions. The instrumental variable approach is based on regional air pollution transport driven by wind direction. We collect reanalysis data on wind conditions from the Japan Meteorological Agency's JRA-55 product. The u- and v-component of wind, i.e. the eastward and northward wind vectors, are reported every six hours (in UTC) on a global grid with a spatial resolution of 1.25° longitude×1.25° latitude, which corresponds to roughly 137.5km×137.5km at the equator.[27] We translate timestamps into local time and aggregate to the daily level. Each city is assigned the inverse distance weighted average of u- and v-vectors at the four grid points located closest to its centroid (median distance = 92.3km). Finally, daily average wind speed and direction are computed from the city-level u- and v-vectors.

Meteorological Conditions. To construct control variables for daily weather conditions we use the ERA5-land product from the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides hourly data on air temperature two meters above the surface, precipitation and dewpoint temperature on a fine grid with 0.1° longitude×0.1° latitude horizontal

[27]We deliberately use data reported on such a coarse spatial grid in order to capture broad wind patterns driving regional air pollution transport instead of very local wind conditions which only affect air quality in a small area. The choice of data follows a suggestion by Tatyana Deryugina which we gratefully acknowledge.

resolution, corresponding to roughly 11km×11km. To construct city×day level variables, we follow the same approach as taken with the wind data, the only difference being that sample cities are assigned the inverse distance weighted average weather conditions from the eight, instead of four, closest grid points (median distance = 10.9km). The variables constructed are daily mean, minimum and maximum temperature, precipitation, and relative humidity.[28]

Wildfire Smoke. The North American west coast frequently experiences severe wildfires generating heavy smoke that strongly increases the concentration of air pollution. Some of the largest cities within our sample are located in this area (the tech clusters in the San Francisco Bay Area and around Seattle). Given recent evidence that exposure to heavy wildfire smoke can trigger avoidance behavior, especially among high-income individuals (Burke et al., 2022), we construct control variables for heavy smoke to make sure the effects we estimate reflect physiological impacts of $PM_{2.5}$ exposure, not behavioral responses to wildfires. The required data is derived from satellite images and provided by the National Oceanographic and Atmospheric Administration's Office of Satellite and Product Operations. It covers the North American continent, is reported at the level of individual smoke plumes, and includes a measure of smoke intensity. We define a city as being affected by a smoke event if the smoke plume overlaps with a 15km radius around its centroid. We aggregate the data to the daily level by summing over the intensity measure of all smoke plumes covering a city on a given day. We define a heavy smoke indicator which is one if the city was covered by a plume of the maximum intensity or if the total daily smoke intensity exceeds a value of twice the maximum intensity, and zero otherwise. This yields 0.3% of all city-by-day observations and 8.9% of all observations with any smoke exposure as heavy smoke days.

Thermal Inversions. In extensions and robustness checks, we use temperature inversions instead of wind direction as an instrument for $PM_{2.5}$ concentration. The required data is obtained from the ECMWF's ERA5 products[29]. Hourly temperature at the surface level as well as several pressure levels is reported on a 0.25° longitude×0.25° latitude grid. We compute the difference between upper air temperature, at the pressure level 25 hPa above the surface, and surface air temperature. Following several recent papers, e.g. Jans et al. (2018a), we then calculate the average temperature difference during local nighttime hours (midnight to 6 am). Cities are assigned the inverse distance weighted average from the four closest grid points. We use the temperature difference as a measure of inversion strength, $inv\ strength_{cd} = \Delta T_{cd}$, to instrument for pollution.

[28] Relative humidity is inferred from mean daily air temperature and dewpoint temperature using the R package `weathermetrics` which uses formulas provided by the US National Weather Service.

[29] We use the products *ERA5 hourly data on single levels from 1979 to present* and *ERA5 hourly data on pressure levels from 1979 to present*

1.4. Research Design

The first part of this section presents our baseline regression model and discusses why endogeneity issues are likely to arise. In the second part, we describe the instrumental variable approach based on wind direction we adopt to address these issues.

Baseline Regression Model. To analyze how short-run variation in local particulate matter concentration affects output and work patterns of professional software developers, we specify a model for outcome y of developers living in city c on day d.

$$y_{c,d} = \beta PM_{c,d} + \mathbf{w}'_{c,d}\gamma + \delta_{R(c)}h_{c,d} + \mu_c + \mu_{R(c),dow(d)} + \mu_{R(c),yr(d),m(d)} + \varepsilon_{c,d} \tag{1.1}$$

Here, $y_{c,d}$ denotes one of the measures of output quantity, quality, or work patterns described in the previous section. We obtain this variable through an auxiliary regression that includes the information available for each individual developer, i.e., her experience in using GitHub and a developer fixed effect. This way we can reduce the computational burden without losing variation in the regressor of interest which is observed at the city-day level. This procedure is common in the literature (e.g. Currie et al., 2015) and asymptotically equivalent to estimating the underlying individual-level regressions (Donald and Lang, 2007). Appendix 1.D provides a more detailed description.

$PM_{c,d}$ is a measure of particulate pollution and varies across cities c and days d. The fixed effect μ_c controls for time-invariant unobserved factors at the city level. Region-year-month fixed effects $\mu_{R(c),yr(d),m(d)}$ capture time-varying productivity shocks common to all developers in a geographic region R. Region×day-of-week fixed effects $\mu_{R(c),dow(d)}$ and an indicator for holidays, $h_{c,d}$, control for fluctuations in work patterns and productivity across days of the week and public holidays. These fluctuations are allowed to vary in intensity across different world regions. $\mathbf{w}_{c,d}$ is a vector of weather variables that can be correlated with air quality and at the same time affect work patterns. It includes a series of indicator variables for daily mean temperature falling into bins defined based on the 5th, 10th, 20th, 35th, 65th, 80th, 90th, and 95th percentiles of the city-specific temperature distributions. The omitted category is temperature falling between the 35th and the 65th percentile. The effects of temperature fluctuations are also allowed to differ across regions R. The vector further contains cubic polynomials of precipitation, relative humidity, and wind speed and a dummy indicating whether the city is affected by heavy wildfire smoke on day d. We weight all regressions by the number of underlying developer observations in each city–day cell and cluster standard errors at the city level.

The coefficient of interest β is estimated from day-to-day variation in city-level pollution and developer output, conditional on average developer output and after netting out other productivity determinants such as weather, seasonality, and region-wide business cycle dynamics.

Since air quality is not assigned randomly to the city-by-day observations, $PM_{c,d}$ may be endogenous in Equation (1.1) due to unobservable factors which co-vary with particulate matter and productivity. Variations in local economic conditions can for instance affect air pollution and developers' output at the same time. Similarly, local events like a football match or the closing of a bridge may impact both traffic and work patterns. The OLS estimate of β would thus likely suffer from omitted variable bias. A second issue is measurement error in developers' pollution exposure, which we cannot observe, but instead have to proxy for by city-level averages. This generates attenuation bias in the OLS estimate. While our regression model includes a wide range of controls to account for sorting into different cities or fluctuations in economic conditions, we still require an exogenous source of variation in local air pollution to address these two concerns.

IV estimation. We address endogeneity in Equation (1.1) by instrumenting local pollution levels with wind direction. This approach was introduced by Deryugina et al. (2019) and is based on the idea that wind direction affects local particulate matter concentration because it is a key driver of pollution transport. Wind blowing from the ocean or less densely populated areas, for instance, carries substantially lower amounts of pollution into the city than wind blowing from more densely populated or industrial areas.

It is important to note that local weather conditions can also depend on wind. For example, wind blowing from the ocean could reduce temperatures. These local conditions could affect labor-leisure trade-offs (Graff Zivin and Neidell, 2014) and thereby the output of developers via channels other than air quality. Therefore, it is important to control for the wide range of weather conditions contained in $\mathbf{w}_{c,d}$ to ensure that the instrument does not violate the exclusion restriction.

The effect of wind direction is certainly not uniform across all cities in our global sample due to differences in geography. In some cases, more pollution might be transported into the city by wind blowing from the east, in other cases, west wind might carry in most pollution. To account for this, we allow the impact of wind on $PM_{2.5}$ to vary. In principle, we could estimate the effect of wind direction separately for each city. In that case, however, the first stage might pick up effects of highly local transport that affects readings at local monitors due to their location relative to the pollution source, but simply redistributes particulate matter within the boundaries of a city. To ensure that the first stage only captures effects of regional pollution transport that changes $PM_{2.5}$ in the whole city, we restrict the effect of wind to vary at a geographically more aggregate level. As suggested by Deryugina et al. (2019), we use a clustering algorithm to assign cities into groups g based on their longitude and latitude. In our baseline specification, we form 50 groups, using hierarchical clustering with a complete-linkage criterion. They are illustrated in Appendix Figure 1.B.2.[30]

We parameterize the pollution-wind relationship by a trigonometric function.[31] By specifying wind direction $\theta_{c,d}$ in radians instead of using many indicators for wind direction bins we can substantially reduce the required number of variables to appropriately model the wind-pollution relationship. The first stage of the IV estimation is as follows.

$$\begin{aligned} PM_{c,d} = \rho_1^g \sin\left(\theta_{c,d}\right) + \rho_2^g \sin\left(\theta_{c,d}/2\right) + \mathbf{w}'_{c,d}\gamma + \delta_{R(c)} h_{c,d} \\ + \mu_c + \mu_{R(c),dow(d)} + \mu_{R(c),yr(d),m(d)} + \varepsilon_{c,d} \end{aligned} \tag{1.2}$$

The coefficients ρ_1^g and ρ_2^g are allowed to vary across city-groups $g \in \{1, 2, ..., 50\}$.

Figure 1.5 illustrates how this trigonometric function can capture the effect of wind direction on $PM_{2.5}$ levels for the city groups represented by Frankfurt, New York City, and Bangalore. The estimated relationships are depicted in blue. They strongly resemble the results we obtain when we instead measure wind direction by eight indicators representing 45° sections of the wind rose, i.e., (0°-45°], (45°-90°], etc. In Appendix Figure 1.B.3 we present the respective plots for all 50 city groups.

We adopt this 2SLS approach for all analyses at the city×date level. In some parts of our analysis, e.g. when exploring effect dynamics or impacts at the monthly level, we use modified versions of this framework that will be presented in the respective sections.

Measures of pollution. Our primary measure of air pollution is daily $PM_{2.5}$ concentration in μg/m^3. As an alternative, we define a binary variable that indicates whether $PM_{2.5}$ is unusually high relative to common levels in city c. More formally, it takes the value one, when the city-day level $PM_{2.5}$ exceeds the city-specific seventy-fifth percentile, $\mathbb{1}\{PM_{c,d} > Q_{.75}\left(PM \mid c\right)\}$. The proposed measure, therefore, captures non-linear effects of pollution and allows these to differ by location.

1.5. Main Results

In this section, we first present results on how $PM_{2.5}$ exposure affects the quantity and quality of output developers produce. Thereafter we show that they use two margins of adjustment,

[30] When using all 220 cities depicted in Figure 1.1 in the clustering algorithm, it forms singletons for some cities which are very distant from their closest neighbor, e.g. Beijing or Salt Lake City. We drop these cities (depicted in light blue) in the main analysis because whenever a city forms its own first-stage group, the first stage might pick up local pollution transport.

[31] We are grateful to Tatyana Deryugina for this suggestion.

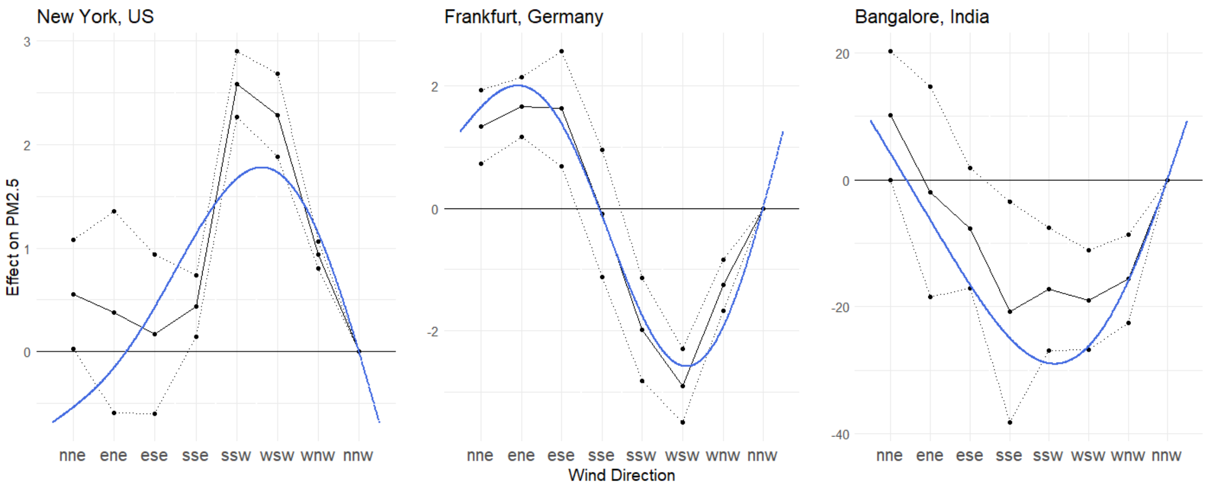


Figure 1.5.: The effect of wind direction on $PM_{2.5}$

Notes: This figure provides a graphical illustration of the first stage. Graphs present estimated coefficients from regressions of $PM_{2.5}$ measured in $\mu g/m^3$ on wind direction. Solid black line: connects estimated coefficients on seven dummies for seven 45° bins of wind direction. The omitted direction is north-north-west, (315°, 360°]. Dashed lines: 95% confidence intervals. Blue line: estimated relationship when wind direction is parameterized as the sine of wind direction in radians and wind direction in radians divided by two. City groups comprise New York and Philadelphia (left), Frankfurt, Nuremberg, Munich, Stuttgart, Karlsruhe, Walldorf, Heidelberg, Bern, Basel, Strasbourg (center), Bangalore, Chennai, Hyderabad (right).

task choice and working hours, to adapt to increases in pollution concentration.

1.5.1. Work Quantity

Columns 1 to 3 of Table 1.2 display 2SLS estimates of the effect of $PM_{2.5}$ exposure on the three primary quantity outcomes—the number of total actions conducted, the number of commits as a measure of individual coding activity, and the number of comments written in discussion fora as a measure of collaborative work. In Panel A, we use $PM_{2.5}$ concentration as regressor and find that an increase by 1 $\mu g/m^3$ causes developers' output, measured by total actions, to fall by 0.0032 or 0.12% of the sample mean. This decline is mainly driven by a reduction in the number of commits, which decreases by 0.0026 or 0.20% of the sample mean. The number of comments is much less affected by air pollution. The point estimate is close to zero and not statistically significant. The first stage F-statistic on the excluded instruments exceeds 100, indicating that the IVs based on wind direction are sufficiently strong. For an increase in ambient $PM_{2.5}$ concentration by one within-city standard deviation (11.8 $\mu g/m^3$), the estimates imply reductions in the number of commits and total actions by 0.030 (2.3%) and 0.038, (1.4%) respectively.

In Panel B, we repeat the analysis, now using the binary variable indicating that $PM_{2.5}$ concentration exceeds the city-specific 75th percentile. The F-statistic is again well above the common threshold for a sufficiently strong first-stage relationship. The 2SLS estimates imply that on a day with relatively high pollution, the number of total actions falls by 0.11 or 4% of the mean value. The number of commits falls by 0.08 or 6.2%. Again, no significant effect on the number of comments is found.

In sum, these results imply that fine particulate matter exposure exerts a negative effect on developer output which is mostly driven by days with relatively poor air quality. The effect of a high-pollution day in Panel B corresponds to an increase in $PM_{2.5}$ concentration by more than 30 $\mu g/m^3$ based on the coefficients in Panel A. A novel finding is the strong effect heterogeneity across different types of work commonly conducted in high-skilled occupations: We observe a highly significant negative impact on individual work on code, but no effect on interactive work.

In Appendix Table 1.A.6, we investigate the effect of $PM_{2.5}$ on further action types which occur less frequently than commits and comments—the number of issues and PRs opened and closed, respectively. Like a commit, opening a PR reflects individual coding work whereas opening/closing issues generally starts/ends a discussion with other users and thus constitutes interactive work. Closing a PR implies decision-making about whether to accept or reject the proposed changes. Consistent with the results discussed above, the number of PRs opened falls significantly in $PM_{2.5}$ concentration, and the relative effect magnitude is similar to the effect on commits. While we find marginally significant effects for closing of PRs, interactive issue events are unaffected by air pollution. Overall, these results confirm the conclusions drawn from Table 1.2.

In Column 4 of Table 1.2 we explore the contribution of the extensive margin to the overall reduction in work quantity. The dependent variable is an indicator for a positive activity level, i.e., $\mathbb{1}\{actions_{id} > 0\}$. For both measures of pollution, we find negative point estimates whose magnitude implies that the extensive margin effect contributes approximately 11-17% to the full reduction in actions. The estimate is statistically significant only for the dummy indicating $PM_{2.5}$ concentration above the 75th percentile. Hence, the extensive margin does explain part of the effect on output, but the intensive margin response is quantitatively much more important. This result is plausible given that our sample of GitHub users likely comprises mostly young and middle-aged adults. They are unlikely to suffer severe health damages from short-run pollution exposure which prevent them from working, especially at moderate levels of concentration, but rather subtle effects on health and cognitive function.

Since we derived our main results by testing eight hypotheses, we also report significance levels that correct for multiple hypothesis testing following the Benjamini-Hochberg procedure in Table 1.2.

In Columns 1 to 3 of Table 1.A.7 we present results from estimating the model in Equation (1.1) by OLS for the three main measures of output quantity to assess the direction and

Table 1.2.: Effect of $PM_{2.5}$ on Work Quantity

	Actions (1)	*Commits* (2)	*Comments* (3)	*Any actions* (4)
Panel A.				
$PM_{2.5}$	-0.0032^{***}	-0.0026^{***}	-0.0005	-0.00013
	(0.0011)	(0.0008)	(0.0005)	(0.00009)
	[0.003]	[0.001]	[0.374]	[0.155]
	{0.006}	{0.002}	{0.374}	{0.208}
First Stage F-Stat.	102.1	102.1	102.1	102.1
% change in Y	-0.12	-0.20	-0.05	-0.04
% of full effect				11.2
Panel B.				
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.1104^{***}	-0.0801^{***}	-0.0169	-0.0068^{***}
	(0.0302)	(0.0159)	(0.0170)	(0.0025)
	[0.0004]	[0.000002]	[0.323]	[0.008]
	{0.0014}	{0.00002}	{0.374}	{0.013}
First Stage F-Stat.	80.5	80.5	80.5	80.5
% change in Y	-4.0	-6.2	-1.8	-1.9
% of full effect				17.1
Observations	353,445	353,445	353,445	353,445
Mean Dep. Var.	2.77	1.29	0.93	0.37

Note: Table presents IV estimates of the parameter β in Equation (1.1). In Panel A, the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Panel B, a binary variable is used instead, which takes a value of one if city×day $PM_{2.5}$ concentration exceeds the city-specific 75th percentile. The first stage specification is given in Equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects, and the temperature controls can vary across world regions *R*. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. Unadjusted p-values and p-values adjusted for multiple hypothesis testing according to the Benjamini-Hochberg procedure are shown in squared and curly brackets, respectively. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

size of the bias. We obtain negative estimates for total actions and commits with both $PM_{2.5}$ in $\mu g/m^3$ and the binary indicator for unusually high pollution levels, but they are significantly different from zero only when using the binary regressor. The results replicate the pattern that effects on commits are larger than on comments. Mirroring a common finding in the literature on short-run impacts of air pollution exposure, all estimates are substantially smaller than the 2SLS results, pointing towards attenuation bias due to measurement error. The ratio of 2SLS-to-OLS estimates ranges between 7 and 13 across specifications.[32] In Table 1.A.8 we re-do the OLS analysis on the extended sample of 220 cities. Results are almost identical for the binary

[32] This is in line with e.g. 2SLS-to-OLS ratios found by Deryugina et al. (2019).

indicator, but the coefficients also attain statistical significance with the continuous regressor.

Effect Magnitude. We conduct three exercises to assess the magnitude and the economic relevance of the estimates. Firstly, we compare the impact of $PM_{2.5}$ concentration above the 75th percentile on developers' output to the effect of another highly relevant environmental shock, exposure to extreme outdoor temperatures.[33] Secondly, we compute elasticities based on the estimated effect of $PM_{2.5}$ on commits and total actions and compare these to elasticities found in previous studies on other occupations. Finally, we leverage the information from Gitcoin to translate the effects into monetary damages.

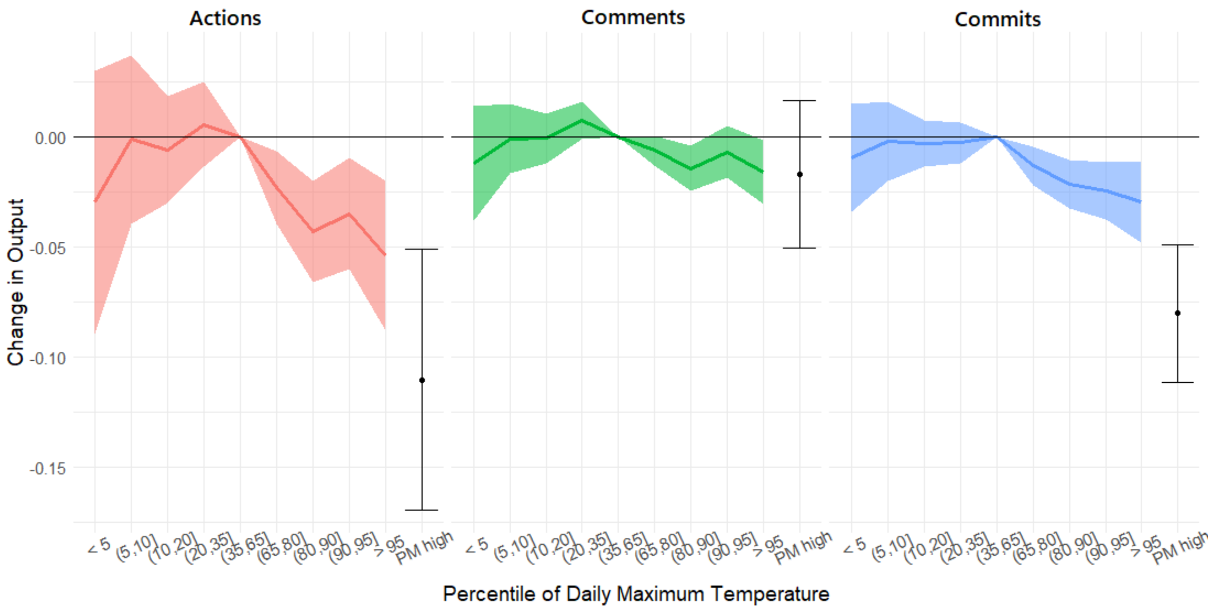
Figure 1.6a reproduces the estimated effects of $PM_{2.5}$ concentration exceeding the 75th percentile on actions, commits, and comments in graphical form (point estimates with 95% confidence intervals displayed in black on the right). In addition, coefficients from OLS regressions of the same outcomes on maximum daily temperature are presented. The regressors of interest are eight dummy variables indicating whether maximum daily temperature falls into a specific percentile range, as displayed on the x-axis. The reference category is a maximum temperature value between the 35th and the 65th percentile. In addition, regressions control for minimum daily temperature, measured in the same way, further weather controls and fixed effects as in Equation (1.1). For all three outcomes, the effects of temperature follow the familiar inverse u-shape: Both unusually cold and unusually hot temperatures have adverse effects, but only the impact of heat is statistically significant.[34] Even though the developers in our samples might work in climate-controlled office buildings, exposure to heat during commuting times or while running other errands might plausibly generate these negative effects. Importantly, for both commits and total actions, the point estimate on the $PM_{2.5}$ dummy is more than twice as large as the point estimate for the highest temperature bin which reflects maximum daily temperature above the 95th percentile.[35] The IV estimates ($PM_{2.5}$) are less precise than the OLS estimates (temperature), but still, even the lower bounds of the 95% confidence intervals on the pollution effects are as large as or even exceeding the point estimates for heat. Hence, the adverse productivity effects of poor air quality exceed those of extreme temperatures, an environmental shock of high relevance given climate change.

Next, we compute elasticities of total actions and commits w.r.t. $PM_{2.5}$ based on the estimates in Panel A of Table 1.2. We obtain elasticities of -.014 and -.025 for actions and commits,

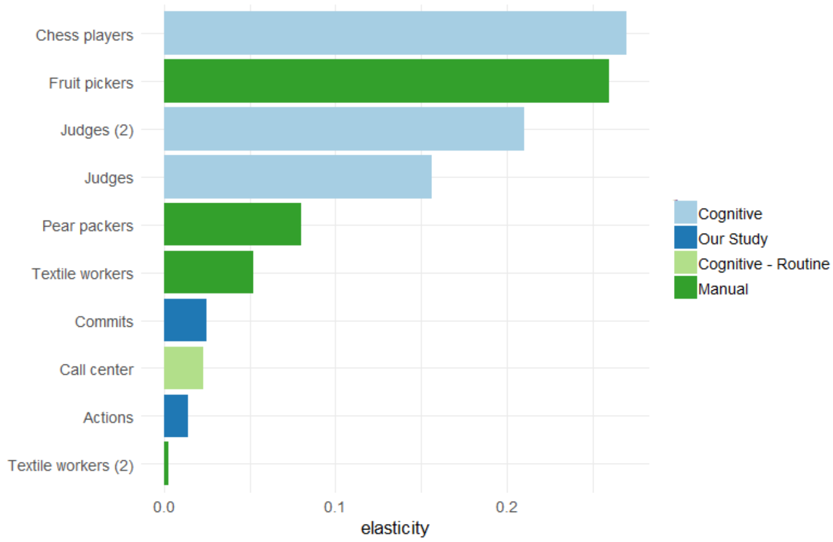
[33] This is motivated by recent findings that, in the U.S., heat exposure exerts adverse effects, e.g., on student performance on high stake exams (Park, 2020), on sentiment among Twitter users (Baylis, 2020), and on mental health (Mullins and White, 2019). Please refer to these papers for more complete overviews of this literature and potential mechanisms.

[34] This is unsurprising given that by analyzing the effects of maximum daily temperature, we can better capture the impact of heat than the effect of cold, and, especially in Europe, not all office buildings are equipped with air conditioning, while heating devices are omnipresent.

[35] Median maximum temperature in this bin is 30.5° C, while the median value in the omitted bin is 16.9° C.



(a) Effects of $PM_{2.5}$ and Heat on Work Quantity



(b) Effects of Air Pollution Across Occupations

Figure 1.6.: Effect Magnitude

Note: Figure 1.6a reproduces the estimated effects of a $PM_{2.5}$ concentration exceeding the city-specific 75th percentile on actions, commits, and comments from Panel B of Table 1.2 in graphical form (point estimates with 95% confidence interval displayed in black on the right). The colored lines represent estimates from an OLS regression of the same outcomes on maximum daily temperature measured by eight dummy variables indicating whether maximum daily temperature falls in a specific percentile range, as displayed on the x-axis. The reference category is a maximum temperature value between the 35th and the 65th percentile. The shaded areas are 95% confidence bands. Control variables are eight corresponding dummies for minimum daily temperature, third-order polynomials for precipitation, wind speed, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects can vary across world regions *R*. Standard errors are clustered at the city level and regressions are weighted by the number of active workers in a city in the current month. Figure 1.6b shows the elasticities of commits and actions with respect to $PM_{2.5}$, based on the estimates in Columns 1 and 2 of Table 1.2. Besides, it presents elasticities of performance with respect to air pollution from other studies, in particular: Künn et al. (forthcoming) (Chess players), Sarmiento (2022) (Judges), Kahn and Li (2020) (Judges (2)), Chang et al. (2019) (Call center agents), He et al. (2019) (Textile Workers (2)), Adhvaryu et al. (2022) (Textile Workers), Chang et al. (2016) (Pear Packers), and Graff Zivin and Neidell (2012) (Fruit pickers).

respectively. These values, along with elasticities of productivity or performance found in previous studies, are depicted in Figure 1.6b.[36] Given that these estimates are derived from very different settings and rely on different approaches (IV vs. OLS estimation, measurements of indoor vs. outdoor pollution), we need to proceed with caution when drawing comparisons between them. However, it stands out very clearly that our estimates are at the lower end of the range of effect sizes found so far. In particular, the effect on developers' output is much smaller than the estimates for judges and chess players, who are also engaged in cognitively demanding tasks. As outlined above, a potential explanation is that chess players and judges face more inflexible settings, namely chess tournaments and court hearings. These circumstances offer no possibility to adapt working hours or the choice of tasks to productivity shocks. This is very different in our setting, and we provide evidence on worker adjustment to an increase in $PM_{2.5}$ in Section 1.5.3. This underscores the importance of our analysis: It might be misleading to draw conclusions on the total economic cost of air pollution based on the estimates for cognitively-demanding tasks in highly inflexible settings because in many high-skilled occupations workers have at least some degree of flexibility in organizing their work day.

Even though productivity effects are small in comparison to other contexts, they might still be economically relevant, given that software development is a high-paying occupation generating large economic value. We use the average monetary value of commits and PRs opened (derived in Section 1.3.3) to translate the estimated negative effects of $PM_{2.5}$ exposure on these two outcomes into changes in output value.[37] For a within-city standard deviation increase in $PM_{2.5}$ (11.8 $\mu g/m^3$) the implied reduction in daily output value amounts to $4.06 per software developer. This is of the same order of magnitude as effects reported by Chang et al. (2016) who find that a 10 $\mu g/m^3$ increase in $PM_{2.5}$ reduces hourly output among pear packers by $0.41, which would imply a damage of $3.28 for a working day of eight hours. On days when $PM_{2.5}$ concentration exceeds the city-specific 75th percentile output value falls by $11.0 relative to days with better air quality. Given that we ignore losses from reductions in task complexity in the calculation (see below), these estimates can be interpreted as a lower bound.

In summary, the impact of air pollution shocks on productivity exceeds the effect of heat. In comparison to other professions, the effect of particulate matter is relatively small, pointing towards an important role of worker adaption in flexible work environments. Economically, the productivity effects are nevertheless relevant, given the high monetary value of software.

[36] Air pollution is measured by $PM_{2.5}$ in all cases except for call center agents and fruit pickers.

[37] As stated in Section 1.3.3 we work with an average monetary value of $112 per commit. In the case of PRs, we do not use the mean value of $354 found in the Gitcoin data because PRs in that sample are larger on average than PRs created in our main analysis sample. Instead, we value PRs with 2.78 × $112 = $311 given that they comprise, on average, 2.78 commits.

Effect Dynamics. The existing literature on air pollution and worker productivity found mixed results on the lagged impact of exposure. He et al. (2019) show evidence for lagged effects of $PM_{2.5}$ and SO_2 on the productivity of textile workers in industrial towns in China, while Künn et al. (forthcoming) find that chess players' performance is unaffected by pollution exposure on the previous days. To explore effect dynamics in our setting, we regress the three measures of output quantity on eleven dummies indicating whether wind was blowing to city c from the high-pollution direction on day d and each of the previous ten days, $\mathit{WDir\ highPM}_{c,d-k}$, where $k \in \{0, 1, \ldots, 10\}$ denotes the lag order. To identify this direction for each city-group g, we run the first stage model with the level of $PM_{2.5}$ concentration as outcome and five dummies for average daily wind direction falling into a specific 60° bin as instruments, interacted with the city-group indicators.

Appendix Table 1.A.9 shows estimated coefficients from the "reduced form" model for total actions, commits, and comments, including just the indicator for the same day, $\mathit{WDir\ highPM}_{c,d}$. The signs and significance of the estimated coefficients are in line with the 2SLS results. The first stage effect, reported at the bottom of the table, implies that wind from a city's high-pollution direction raises $PM_{2.5}$ concentration on average by 3.7 µg/m^3 relative to days where wind arrives from another direction. While this approach is much less flexible than our main 2SLS model, it captures the underlying idea in a single variable and thus allows us to easily analyze effect dynamics by including lags. Since we are mostly interested in qualitative results – does lagged exposure reduce productivity or does it induce developers to work more in order to catch up – we opt for this approach.

The distributed lag model we use to explore effect dynamics includes the same covariates as our main contemporaneous model plus ten lags of weather conditions and is given by:

$$y_{c,d} = \sum_{k=0}^{10} \beta_k \mathit{WDir\ highPM}_{c,d-k} + \sum_{k=0}^{10} \mathbf{w'}_{c,d-k}\gamma_k + \mu_c + \mu_{R(c),yr(d),m(d)} + \mu_{R(c),dow(d)} + \delta_R h_{c,d} + \varepsilon_{c,d} \quad (1.3)$$

From the corresponding estimates $\widehat{\beta_k}$, we compute the cumulative effect of exposure to wind from the high-pollution direction for s consecutive days, $\sum_{k=0}^{s} \hat{\beta}_k$ for $s = 0, 1, \ldots, 10$, which we plot in Figure 1.7.

For both total actions and commits, same-day exposure to pollution generates negative and significant effects. The cumulative effect magnitude grows monotonically up to the third lag. The current day effect of $\mathit{WDir\ highPM}_{c,d}$ is estimated to be -.018 for total actions and -.007 for commits. After four consecutive days of wind from the high-pollution direction, the cumulative effect is -.037 and -.022, respectively. For more prolonged exposure the point estimate of the cumulative effect remains rather constant but becomes noisier, likely due to serial cor-

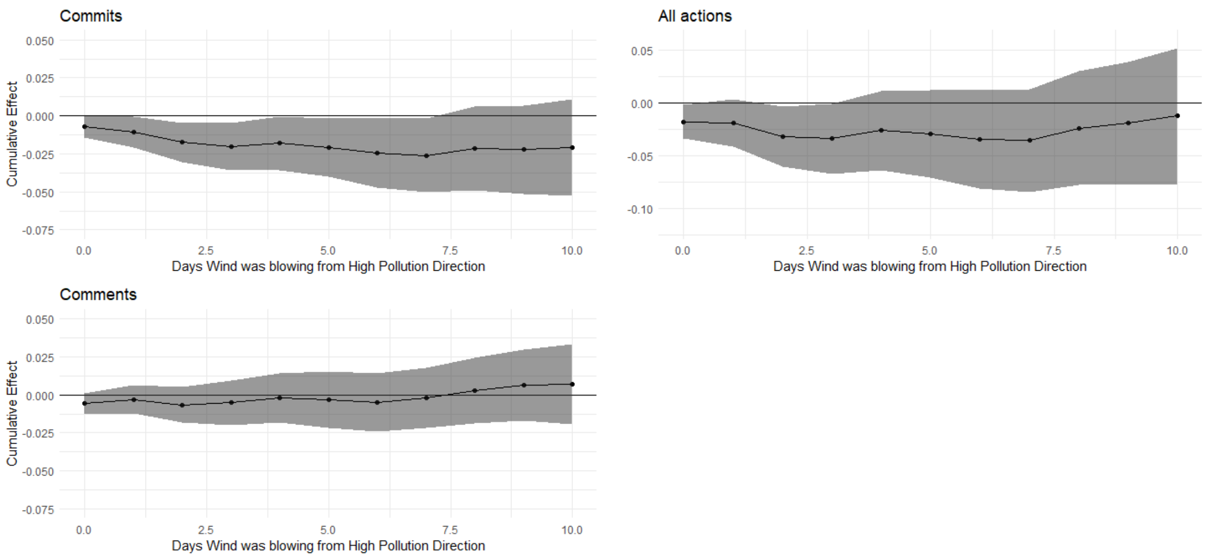


Figure 1.7.: Effect Dynamics: Work Quantity

Note: The plots depict estimates of the cumulative effect of wind blowing from the high-pollution direction on three measures of work quantity. Effects are derived from a distributed lag model and given by $\sum_{k=0}^{s} \hat{\beta}_k$ for $s = 0, 1, \ldots, 10$. The x-axis denotes the number of days over which the cumulative effect is computed. Shaded areas represent 95% confidence intervals. Regressions control for city, Region-by-day-of-week, and Region-by-year-by-month fixed effects, a holiday indicator, and weather controls for the current day and ten lags (third-order polynomials in mean daily temperature, precipitation, relative humidity, and wind speed). Regressions are weighted by the number of active workers in a city during the current month and standard errors are clustered at the city level.

relation in the wind direction variable. For commits, there is no decline in the cumulative effect magnitude at higher lags, i.e., no evidence that developers completely make up for the output loss within the first ten days after exposure. After ten days, the point estimates for commits and total actions are still negative and as large as or even larger than the point estimate for same-day exposure. For comments, the cumulative effect always remains close to zero throughout the full time window.

In sum, pollution exposure generates an adverse effect on current-day output and, to a smaller extent, also reduces productivity on the following three days. Compared to health impacts, the productivity effects are rather immediate.[38]

1.5.2. Work Quality

So far, we have shown that exposure to $PM_{2.5}$ reduces the quantity of output developers produce per day. In high-skill jobs, output quality is of major relevance and might also be affected by pollution shocks.

Table 1.3 displays 2SLS estimates of the effect of $PM_{2.5}$ on two measures of work quality. The first is the share of all PRs a user opened on a given day that is later merged, i.e., accepted. PR

[38]Barwick et al. (2018) for instance find that $PM_{2.5}$ exposure raises medical expenditures up to 90 days post exposure.

rejections suggest issues with code quality or style, indicating low work quality. The second is the share of commits made by a user on a given day that were later reverted. Commit reversals point toward severe errors that cannot easily be corrected in follow-on commits, i.e., major issues with the work quality. Sample sizes are reduced relative to the results on output quantity, because the outcomes are only defined for city×day observations with any PRs opened and any commits, respectively. Moreover, information on commit reversals is from the GHArchive data which is only available from 2015 onward.

We find small, insignificant point estimates for both outcomes when using $PM_{2.5}$ in μg/m^3 as regressor. PRs opened on days when $PM_{2.5}$ concentration exceeds the 75th percentile, are 0.8 p.p. more likely to get accepted. This represents an increase of 1.3% relative to the mean, i.e., a small improvement in work quality. Estimates remain insignificant for the share of commits that are reverted, but the negative sign is also in line with minor reductions in error frequency.

Table 1.3.: Effect of $PM_{2.5}$ on Work Quality

	Share PRs merged	*Share Commits reverted*
	(1)	(2)
Panel A.		
$PM_{2.5}$	0.00013	−0.0000004
	(0.00019)	(0.00001)
First Stage F-Stat.	56.0	78.8
% change in Y	0.19	- 0.02
Panel B.		
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	0.0083**	−0.00015
	(0.00399)	(0.00022)
First Stage F-Stat.	42.9	61.1
% change in Y	1.3	-8.2
Observations	135,433	215,728
Mean Dep. Var.	0.665	0.002

Note: The table presents IV estimates of the parameter β in Equation (1.1). The outcome in Column (1) is defined as the share of all PRs opened by a developer on a given day that later gets accepted. The outcome in Column (2) is defined as the share of all commits made by a developer on a given day that later get reverted, i.e., undone (see Section 1.3 for details). In Panel A, the regressor of interest is $PM_{2.5}$ concentration measured in μg/m^3. In Panel B, a binary variable is used instead, which takes a value of one if city×day $PM_{2.5}$ concentration exceeds the city-specific 75th percentile. The first stage specification is given in Equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and the temperature controls can vary across world regions *R*. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

The null effect on quality contrasts with findings by Archsmith et al. (2018) who show that baseball umpires conduct more errors when exposed to higher pollution levels. In the next section, we present evidence that developers change their work patterns when exposed to high levels of pollution. This form of adaptation might explain why effects on output quantity are relatively modest and quality is unaffected in this flexible high-skilled setting.

1.5.3. Worker Adjustment

In this section, we investigate whether work patterns change in response to increases in particulate pollution. We consider two potential margins of adjustment, task choice and working hours.

Switching to Easy Tasks. We analyze whether developers switch towards easier tasks when exposed to higher levels of pollution, for both activities related to issues, i.e., interactive tasks, and activities related to pull requests, i.e., coding and review tasks. Table 1.4 presents estimates of the impact of pollution on the share of issue events completed that refer to an easy issue (Column (1)). As this outcome is only defined for city×day observations with non-zero issue events (issue opened, closed, or reopened, or a comment written on an issue), the number of observations is reduced. We find that the share of events referring to easy issues increases if $PM_{2.5}$ concentration rises. In terms of magnitude, an increase in pollution by one within-city standard deviation raises the share by 2.6%. On days when fine particulate matter levels exceed the city-specific 75th percentile, the variable even increases by 4.9% of the mean, or 0.32 percentage points.

In the second column, we analyze how the denominator of the share changes. Consistent with earlier results on the impact of $PM_{2.5}$ on interactive tasks, we find no statistically significant effect on the number of issue events. Thus, changes in the share of easy issue events are not driven by changes in the denominator, but by a switch towards more easy issues for a relatively constant activity level with respect to issue events. When hit by a pollution-induced productivity shock, developers seem to exploit the fact that certain issue labels provide a prominent signal of issue complexity to focus on easier tasks.

This finding is corroborated when we consider the complexity of PRs in Columns (3) to (5). We consider three PR characteristics: lines of code added, lines of code deleted, and number of files changed, averaged across all PRs a developer worked on a given day.[39] While there can be very difficult tasks that involve a lot of thinking but require only small changes in the code, we believe that these variables provide reasonable proxies of PR complexity. Fixing a severe

[39] These variables are based on GHArchive data, whereas work quantity results used GHTorrent data. GHArchive data is only available from 2015 onward. In Appendix Table 1.A.10 we show that using data on pull requests from GHArchive we can replicate the results presented in Table 1.A.6 for PRs opened or closed measured in the GHTorrent data.

Table 1.4.: Effect of $PM_{2.5}$ on Task Complexity

	Share Easy Issue Events (1)	*Issue Events* (2)	*Lines added per PR* (3)	*Lines deleted per PR* (4)	*Files changed per PR* (5)
Panel A.					
$PM_{2.5}$	0.00015^{**}	-0.00046	-0.0013	-0.0013	-0.0010^{**}
	(0.00007)	(0.00041)	(0.0008)	(0.0010)	(0.0004)
	[0.036]	[0.264]	[0.112]	[0.202]	[0.017]
First Stage F-Stat.	86.3	102.1	62.1	62.1	62.1
% change in Y	0.2	-0.5	-0.1	-0.1	-0.1
Panel B.					
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	0.0032^{**}	-0.0219	-0.0404^{**}	-0.0196	-0.0244^{**}
	(0.0015)	(0.0134)	(0.0181)	(0.0214)	(0.0098)
	[0.028]	[0.105]	[0.027]	[0.362]	[0.014]
First Stage F-Stat.	66.1	80.5	45.6	45.6	45.6
% change in Y	4.9	-2.4	-4.0	-2.0	-2.4
Observations	250,376	353,445	164,883	164,883	164,883
Mean Dep. Var.	0.067	0.90			

Note: The table presents IV estimates of the parameter β in Equation (1.1). The outcome in Column (2) is defined as the sum of actions referring to issues, i.e., the number of issues opened, closed, reopened, and the number of issue comments written. The outcome in Column (1) is defined as the ratio of the number of these activities which refer to an issue classified as *easy* based on issue labels (see Section 1.3 for details) and the total number of issue events. In Columns (3) to (5), outcomes are defined as the average number of lines of code files changed, number of new lines of code added, and number of lines of code deleted across all PRs a developer opened, closed, or commented on. Inverse hyperbolic sine transformations are applied to these outcomes. In Panel A, the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Panel B, a binary variable is used instead, which takes a value of one if city×day $PM_{2.5}$ concentration exceeds the city- specific 75th percentile. The first stage specification is given in Equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and temperature controls can vary across world regions R. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. *p<0.1; **p<0.05; ***p<0.01

bug for instance likely requires changes in different parts of the source code, which implies a larger number of files changed. Results presented in Table 1.C.3 indicate that PRs with more lines of code added and files changed are rewarded higher payments on Gitcoin, validating the use of these variables as complexity metrics. Similarly, reviewing a PR is more demanding when it contains large changes across different files. The characteristics we use to measure complexity are prominently displayed when opening a PR on GitHub, such that reviewers can easily assess them and jugde their difficulty level.

We apply the inverse hyperbolic sine transformation to the outcome variables such that coefficients approximate percentage changes. The sample size is reduced relative to previous tables because the outcomes are defined only for city×day observations with a positive number of PRs opened, closed, or commented on. Point estimates are negative across all three outcomes and the two distinct measures of air pollution. On days with unusually high $PM_{2.5}$ levels, the number of files changed falls by 2.4%, while the number of lines added drops by 4%. The effect on the number of lines deleted is also negative, but only half as large and not significantly different from zero. This pattern is plausible since tasks related to deleting code, e.g., cleaning or polishing a file or dropping a deprecated or redundant part, are often easier than creating new code. The pattern is similar for the continuous regressor, but the effect on new lines added is not significant at conventional levels, indicating that developers move towards less complex tasks mostly in response to large productivity shocks on high-pollution days.

In sum, the results imply that, on top of the overall reduction in the number of actions completed, developers switch towards less complex tasks when exposed to high levels of $PM_{2.5}$. Thus, the estimates of monetary effects of air pollution exposure presented above provide a lower bound, given that we found in Section 1.3.3 that less complex pull requests and those addressing issues labeled as *easy* are rewarded lower payments, even when holding the number of commits constant.

This form of adjustment might also explain why the magnitude of effects on output quantity is relatively small in comparison to results found for other occupations, and why work quality is not adversely affected. We next present some evidence that switching to easier tasks is indeed an adaptation strategy to surges in particulate matter exposure that allows developers to prevent large declines in work quantity. Table 1.5 presents results from a heterogeneity analysis based on developer characteristics. Specifically, we combine tenure (time since registration on GitHub) and the number of followers at the point in time the developer enters our sample into an index that represents their experience and popularity.[40] We split the sample developers into terciles based on experience and run the IV regression at the developer×day level. Panels A and B present results estimated separately for the bottom and the top tercile, respectively.[41]

The first three columns display estimated effects of $PM_{2.5}$ on the three primary measures of work quantity. Point estimates for total actions and commits are negative in both samples, but larger in absolute terms as well as relative to the sample means for the most experienced developers. Effects on comments are not significantly different from zero in either sample. For the adaptation variables examined in the last three columns, the pattern is reversed: While the direction of the effects is again the same in both samples, effects are now stronger among the less experienced developers and not significantly different from zero in the upper tercile.

These results confirm that switching to easier tasks is a form of adjustment to pollution-induced productivity shocks among highly-skilled workers. A potential reason why the least

[40] The index is computed as the average of tenure and the number of followers, after standardizing both variables.

[41] Median tenure (number of followers) is 1.4 years (11) in the bottom tercile and 5.7 years (31) in the upper tercile.

Table 1.5.: Effect Heterogeneity by Experience

	Actions (1)	*Commits* (2)	*Comments* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A: Bottom Tercile of Experience						
$PM_{2.5}$	-0.0033^{*}	-0.0026^{**}	-0.0008	0.00035^{**}	-0.0022	-0.0019^{**}
	(0.0017)	(0.0011)	(0.0008)	(0.00014)	(0.0015)	(0.0009)
First Stage F-Stat.	3187	3187	3187	586	211	211
% change in Y	-0.12	-0.19	-0.10	0.47	-0.22	-0.19
Observations	4,774,247	4,774,247	4,774,247	900,318	327,297	327,297
Mean Dep. Var.	2.48	1.24	0.77	0.074		
Panel B: Upper Tercile of Experience						
$PM_{2.5}$	-0.0062^{***}	-0.0044^{***}	-0.0009	0.0001	-0.0015	-0.0003
	(0.0023)	(0.0014)	(0.0010)	(0.0001)	(0.0023)	(0.0013)
First Stage F-Stat.	3010	3010	3010	756	222	222
% change in Y	-0.18	-0.30	-0.06	0.17	-0.15	-0.03
Observations	4,387,377	4,387,377	4,387,377	1,105,410	324,527	324,527
Mean Dep. Var.	3.16	1.37	1.15	0.061		

Note: The table presents IV estimates of the parameter β in equation (1.1). Inverse hyperbolic sine transformations are applied to outcomes in Columns 5 and 6. The regressor of interest is $PM_{2.5}$ concentration measured in city-specific standard deviations. The first stage specification is given in equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as developer, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and temperature controls can vary across world regions *R*. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

experienced developers show a stronger adjustment response might be that they have the largest incentive to keep up a high activity level because the number of actions performed per day in public repositories is visualized on a user's GitHub profile and might be an important signal to peers or potential employers. Work complexity on the other hand is less easily observable.

Working Hours. A second potential adjustment margin available in flexible work environments is a change in working hours. We start by analyzing whether developers expand or reduce activity in the evening in response to increasing pollution exposure. Table 1.6 presents the estimated effects of $PM_{2.5}$ on the timestamp of the last action performed by a developer on a given day (in minutes) and on the share of total actions conducted after 6 pm.[42]

[42]These outcomes are only defined for developer×day observations with at least one action, which explains the reduction in sample sizes.

Table 1.6.: Working Hours

	Time of Last Action (minutes)	*Share of Actions after 6 pm*
	(1)	(2)
Panel A.		
$PM_{2.5}$	-0.226^{***}	-0.0001
	(0.085)	(0.0001)
	[0.009]	[.147]
First Stage F-Stat.	96.1	96.1
Panel B.		
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-4.162^{**}	-0.002
	(2.061)	(0.003)
	[0.045]	[.368]
First Stage F-Stat.	73.9	73.9
Observations	302,575	302,575

The table presents IV estimates of the parameter β in Equation (1.1). In Panel A, the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Panel B, a binary variable is used instead, which takes a value of one if city×day $PM_{2.5}$ concentration exceeds the city-specific 75th percentile. The outcome variables are the timestamp of the last action conducted by a developer on a given day in minutes in Column (1), and the share of all activities conducted after 6 pm in Column (2). Estimates are based on all developer×date observations with at least one recorded action in Column (1) and at least two recorded actions in Column (2). The first stage specification is given in Equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and the temperature controls can vary across world regions *R*. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. *p<0.1; **p<0.05; ***p<0.01

We find that developers on average end the work day 0.23 minutes earlier in response to an increase in $PM_{2.5}$ concentration by 1 $\mu g/m^3$. On days with $PM_{2.5}$ levels above the city-specific 75th percentile, the work day ends 4.2 minutes earlier than on days with better air quality. To put this into perspective, we approximate the average time spent per action as

$$\frac{1}{N}\sum_i \sum_d \Big(timestamp(last\ action_{i,d}) - timestamp(first\ action_{i,d}) \Big) / \Big(number\ of\ actions_{i,d} - 1 \Big)$$

where N denotes the number of developer×date observations with at least two actions. This yields an average of 72.5 minutes spent per action.[43] Point estimates for the share of actions conducted in the evening are also negative, but small and not statistically significant. Subtle effects of pollution might make developers feel unproductive, inducing them to end their work activity earlier on high-pollution days due to reduced opportunity cost of leisure time. If $PM_{2.5}$

[43]Importantly, this computation gives the time input including breaks, as we cannot disentangle time spent working on activities and breaks.

exposure, e.g., triggers headaches or fatigue, developers might experience this as an off day and decide to reallocate work to days when they perform better. In many jobs, knowledge workers are very flexible in when and where they want to work. Thus, shifting work intertemporally from low productivity days to the weekend, a period with relatively low activity levels and thus scope for compensation (see Figure 1.2), might be an important adjustment margin in these settings. To investigate this, we estimate the effect of $PM_{2.5}$ exposure during the first half of the workweek on output produced on the weekend. This analysis is conducted at the developer×week level, using the following, slightly modified, regression model.

$$\begin{aligned} y_{i,c,w}^{weekend} = \beta PM_{c,w}^{Mo\text{-}We} + \mu_i + \mathbf{x}_{i,t}'\pi + \mathbf{w}'^{weekend}_{c,w}\gamma + \mathbf{w}'^{Mo\text{-}We}_{c,w}\alpha + \\ \delta_{R(c)}h_{c,w} + \mu_{R(c),yr(w),q(w)} + \mathbf{z}'_{c,w}\varphi + \varepsilon_{i,c,w} \end{aligned} \tag{1.4}$$

$y_{i,c,w}^{weekend}$ denotes the sum of actions conducted by developer i living in city c on the weekend of week w. $PM_{c,w}^{Mo\text{-}We}$ is a measure of $PM_{2.5}$ concentration in city c between Monday and Wednesday of week w. Specifically, we consider average $PM_{2.5}$ concentration or the number of days with $PM_{2.5}$ concentration exceeding the city-specific 75th percentile. Due to the finding that exposure to pollution exerts negative effects on output not only on the same day, but also over the next two to three days, we focus on $PM_{2.5}$ during the first half of the workweek to make sure that we pick up developers' *behavioral* adjustment to a productivity shock during the workweek, and do not confound it with *physiological* effects.

Pollution is instrumented by the same variables as described in Equation (1.2), with the only difference that wind direction $\theta_{c,w}$ is averaged between Monday and Wednesday. To account for auto-correlation in the instruments, we add the vector $\mathbf{z}_{c,w}$ to the model, which includes the instrumental variables measured on the weekend and on Thursday to Friday. This ensures that we do not pick up the effects of wind direction-induced changes in pollution on the weekend itself or the days immediately before. The model further includes a developer fixed effect μ_i, a region-by-year-by-quarter fixed effect, $\mu_{R(c),yr(w),q(w)}$, the number of public holidays during the workweek, $h_{c,w}$, a vector $\mathbf{x}_{i,t}'$ of bin variables capturing the developer's tenure on GitHub, and two sets of weather controls, $\mathbf{w}_{c,w}^{Mo\text{-}We}$ and $\mathbf{w}_{c,w}^{weekend}$, covering the exposure period and the weekend, respectively.[44]

Table 1.7 presents estimates of coefficient β. Results in Columns (1), (3), and (5) indicate that developers produce significantly more output on weekends if they were exposed to unusually high levels of $PM_{2.5}$ between Monday to Wednesday of the same week (Panel B). In

[44] $\mathbf{w}_{c,w}^{Mo\text{-}We}$ and $\mathbf{w}_{c,w}^{weekend}$ both comprise the number of days with heavy wildfire smoke exposure as well as third-order polynomials in average precipitation, relative humidity, and wind speed during the respective period. $\mathbf{w}_{c,w}^{Mo\text{-}We}$ further includes eight variables counting the number of days on which daily mean temperature is falling into the temperature bins described above. $\mathbf{w}_{c,w}^{weekend}$ includes a third-order polynomial in average temperature.

Table 1.7.: Effect of $PM_{2.5}$ in the First Half of the Workweek on Weekend Work

	Actions		*Commits*		*Comments*	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A.						
$PM_{2.5}$	0.0043	0.0133**	0.0017	0.0062**	0.0016	0.0037**
	(0.0030)	(0.0055)	(0.0017)	(0.0030)	(0.0012)	(0.0019)
	[0.158]	[0.016]	[0.308]	[0.041]	[0.158]	[0.048]
First Stage F-Stat.	1406	1147	1406	1147	1406	1147
Panel B.						
High $PM_{2.5}$ Days	0.0630***	0.1362***	0.0268*	0.0646***	0.0213**	0.0341***
	(0.0225)	(0.0321)	(0.0136)	(0.0207)	(0.0084)	(0.0107)
	[0.006]	[.00004]	[0.051]	[0.002]	[0.013]	[0.002]
First Stage F-Stat.	1455	801	1455	801	1455	801
Observations	2,011,797	1,352,191	2,011,797	1,352,191	2,011,797	1,352,191
Weeks	all	only low PM weekends	all	only low PM weekends	all	only low PM weekends

Note: The table presents IV estimates of the parameter β in equation 1.4. Outcomes are the sum of all actions, commits, and comments made on the weekend, respectively. In Panel A, the regressor of interest is average $PM_{2.5}$ concentration between Monday and Wednesday. In Panel B, the count of days on which the city×day $PM_{2.5}$ concentration exceeds the city- specific 75th percentile during this period is used instead. The first stage specification is given in equation 1.2. Regressions control for developer and region-by-year-by-quarter fixed effects, the number of public holidays during the workweek, and the leads of the instrumental variables for both the weekend and the period from Thursday to Friday. Further covariates are the number of days with heavy wildfire smoke, and third-order polynomials in average wind speed, precipitation, and relative humidity during both the weekend and the period between Monday and Wednesday. Temperature controls are included in the form of eight bin variables for the period Monday to Wednesday, and in the form of a third-order polynomial for the weekend, and are allowed to vary across regions R. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. *p<0.1; **p<0.05; ***p<0.01

terms of magnitude, one additional day with $PM_{2.5}$ concentration exceeding the city-specific 75th percentile causes an increase in total actions on the weekend by 0.063 or 2.1% of the mean. Effects are positive and significant for both commits and comments, amounting to 1.6% and 2.8% of the mean values, respectively. When we consider average $PM_{2.5}$ concentration between Monday and Wednesday instead, we find positive point estimates, but these are not significantly different from zero. The compensation thus seems to be most relevant after high-pollution days, which is consistent with the finding that unusually high pollution levels result in disproportionately large declines in output quantity on the day of exposure.

In Columns (2), (4), and (6), we repeat the same analysis, but using only developer×week observations with low pollution levels on the weekend, i.e., levels below the city-specific 75th percentile on both days. We find substantially larger coefficients, indicating that developers reallocate work from low to high productivity periods, i.e., weekends without air pollution-induced productivity shocks.

To put these effects into perspective, we can compare the effect magnitudes with the estimates in Table 1.2 depicting the reductions in daily output due to same-day $PM_{2.5}$ exposure. Additional work on the weekend makes up for 33% and 57% of the reduction in commits and total actions due to $PM_{2.5}$ exceeding the 75th percentile, respectively.[45]

To check that the estimates in Table 1.7 do not pick up any effects of unobservable confounders, but indeed reflect a behavioral response of developers to pollution-induced productivity shocks, we conduct a falsification test. We shift both the weekend and the exposure period forward by four days. The placebo weekend comprises Wednesday and Thursday and the placebo exposure period ranges from Friday to Sunday of the week before. Since activity levels are low on weekends, productivity shocks on these days should not induce compensation during the following week. Moreover, activity is already high on Wednesday and Thursday, such that there is not much scope for additional work. Hence, we expect no significant effects of $PM_{2.5}$ exposure. Appendix Table 1.A.11 presents the results which confirm this hypothesis. Effects are neither significant in the full sample, nor when considering only placebo weekends with low pollution levels.

In summary, we find that developers work more on weekends to catch up on coding tasks not completed due to pollution-induced productivity declines during the workweek.[46] The opportunity for compensation might allow them to end work early on high-pollution work days. This reallocation option could thus also contribute to the absence of effects on work quality in this setting. If developers can end work when their health or cognitive capacity deteriorates and they face an increased risk to commit errors, this will mitigate impacts of pollution on work quality. At the same time, sacrificing leisure time on the weekend, when it is likely most valuable, implies a welfare cost and potentially adverse effects on the work-life-balance.

Overall, worker adaptation likely plays an important role in explaining the modest effects of $PM_{2.5}$ on output. By focusing on easier tasks and reallocating work from high-pollution, low-productivity to low-pollution, high-productivity periods, developers alleviate the impact of the environmental shock.

[45] In the case of comments, we find positive effects on weekend activity, even though there are no significant reductions in comments due to pollution exposure. When developers work on the weekend because they are behind on coding tasks, they might decide to also conduct some interactive actions given that they are active on GitHub anyways, even though there was no negative effect of pollution exposure in this domain. This might explain the positive effect for comments and the large coefficient for total actions relative to the direct effect.

[46] This result is similar to the finding by Hoffmann and Rud (2022) that workers in Mexico City reallocate labor supply across days in response to changes in $PM_{2.5}$. However, the authors find strong extensive margin effects and show that reallocation likely serves as a strategy to *avoid* pollution exposure and its adverse health impacts. In our setting focused on a global sample of high-skill tech workers reallocation is likely rather a response to low productivity during high-pollution periods.

1.6. Heterogeneity and Further Results

Our main results are based on a linear measure of $PM_{2.5}$ concentration and an indicator for unusually high levels. In this section, we exploit the large variation in air quality in our international sample to investigate whether effects on output arise across the full range of concentrations, and how they vary in intensity. Furthermore, we analyze effect heterogeneity based on location characteristics, repeat our analysis at the monthly level, and conduct several robustness checks.

Non-Linearity. To analyze the shape of the dose-response function between $PM_{2.5}$ and output quantity, we replace $PM_{c,d}$ in equation (1.1) with a series of dummy variables indicating whether $PM_{2.5}$ concentration falls into a specific bin.[47] We estimate the model by OLS on the extended sample of 220 cities. Since we cannot rely on exogenous variation in air quality due to wind direction in this analysis, we opt for a more conservative specification with stricter fixed effects for region×date and city×month. These absorb (i) region-wide shocks to developer output on a given date that might be correlated with $PM_{2.5}$ concentration and (ii) seasonal fluctuations in activity and air quality which are allowed to vary across cities.[48] Given the finding that the OLS results underestimate the true effects, we need to bear in mind that all results should be interpreted as reflecting lower bounds.

Figure 1.8 displays estimated effects of the $PM_{2.5}$ bin variables on actions and commits. Estimated coefficients reflect the impact of moving from a $PM_{2.5}$ concentration between 16 and 22 $\mu g/m^3$ to the respective bin. The x-axis measures the average concentration within the bins. The baseline bin is chosen to ensure that for each city some observations fall into this range. For perspective, it falls below the EU limit value for $PM_{2.5}$ during the sample period (25 $\mu g/m^3$), but above the EPA annual standard (12 $\mu g/m^3$). We find significant negative effects starting at a concentration of approximately 75 $\mu g/m^3$, but no significant differences in the outcomes for concentrations between the reference bin and 60 $\mu g/m^3$. Being exposed to a $PM_{2.5}$ level below 5 $\mu g/m^3$ has a significant positive impact on both total actions (point estimate = 0.038, p-value = 0.061) and commits (point estimate = 0.020, p-value = 0.031). This implies that even in cities with low to moderate levels of $PM_{2.5}$, further improvements in air quality will generate positive effects on worker productivity.

For both outcomes, the average slope of the function is larger than the estimate found in the linear OLS specification, especially at low $PM_{2.5}$ concentrations. To zoom in on the differ-

[47] Bins are defined for ≤ 5, (5-6], (6-8], (8-10], (10-13], (13,16], (16,22], (22,25], (25,28], (28,31], (31,35], (35,40], (40,50], (50,60], (60, 70], (70, 85], (85,100], (100, 160] and > 160, all in $\mu g/m^3$, with (16-22] as reference bin.

[48] OLS results from this model with either $PM_{2.5}$ in $\mu g/m^3$, or the binary indicator for $PM_{2.5}$ above the city-specific 75th percentile as regressors are presented Columns 4 to 6 of Tables 1.A.7 and 1.A.8 for the main sample and the extended sample, respectively.

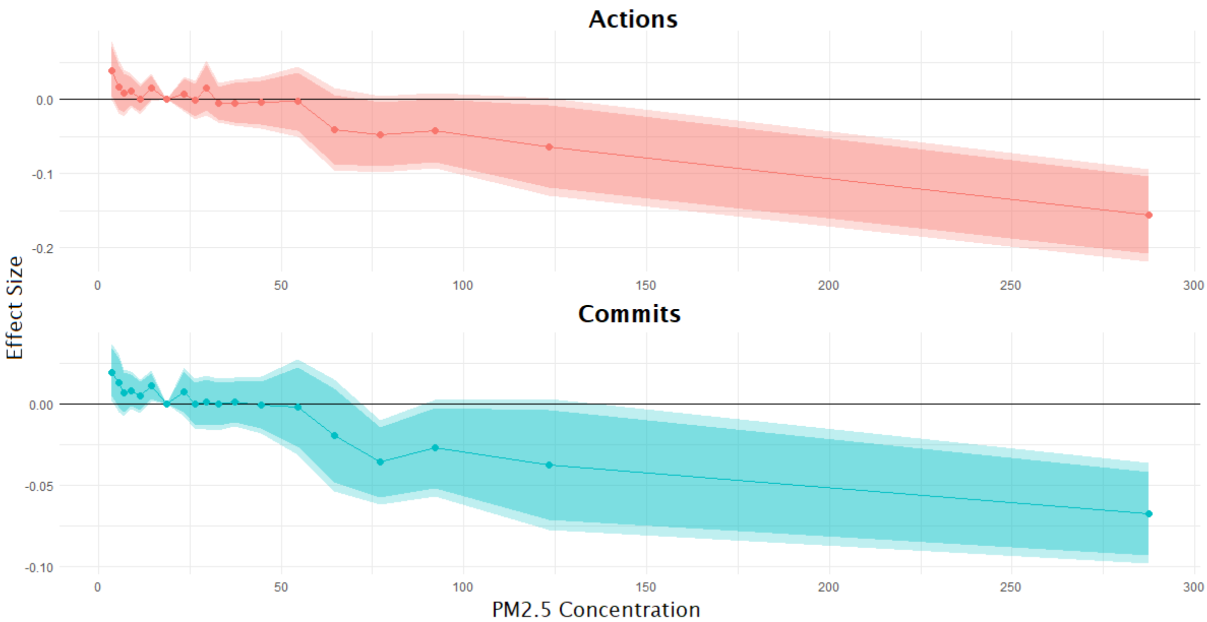


Figure 1.8.: Non-linear effects of $PM_{2.5}$ on Work Quantity (OLS estimates)

Note: Plot depicts point estimates on different bins of $PM_{2.5}$ concentrations from an OLS regressions of total actions (left) and commits (right), respectively, on indicators for each bin. Covariates: Weather and holiday controls as in Equation 1.1, region×date and city×month fixed effects. X-axis: Average $PM_{2.5}$ concentration in each bin in $\mu g/m^3$. Shaded areas indicate 95%- and 90%-confidence intervals.

ent parts of the function, we split the sample into terciles based on cities' average pollution concentration. For each subsample we define seven bin variables for $PM_{2.5}$ such that each bin includes the same number of city×date observations. The reference category is given by the lowest bin. Figure 1.9 presents the results, focusing on total actions for the sake of exposition.

In the subsample of cities in the bottom tercile, mean $PM_{2.5}$ concentration ranges between 5.1 $\mu g/m^3$ and 8.6 $\mu g/m^3$. Most of the distribution falls below current regulatory thresholds.[49] Moving from the lowest bin to higher concentrations generates significant negative effects on output. The implied slope is -0.0044, i.e., much steeper than the estimate from the linear specification on the full sample, and even exceeds the size of the 2SLS estimate. In the middle tercile, by contrast, the estimates imply a flat slope. Average $PM_{2.5}$ concentration in this sample ranges between 8.7 $\mu g/m^3$ and 13.0 $\mu g/m^3$, and it includes mostly cities in the European Union and the US. Among the most polluted cities in the upper tercile we again find a negative slope, but less steep than in the low pollution subsample. This subsample includes most Asian and Eastern European cities, as well as some cities in Western Europe, with mean concentration ranging between 13 $\mu g/m^3$ and 133 $\mu g/m^3$.

We are unable to pin down what drives the differences in the dose-response function across samples. They might arise because of differences in the extent of measurement error or omit-

[49]This subsample includes most cities in Australia, New Zealand, Scandinavia, and Canada and more than half the cities in the US.

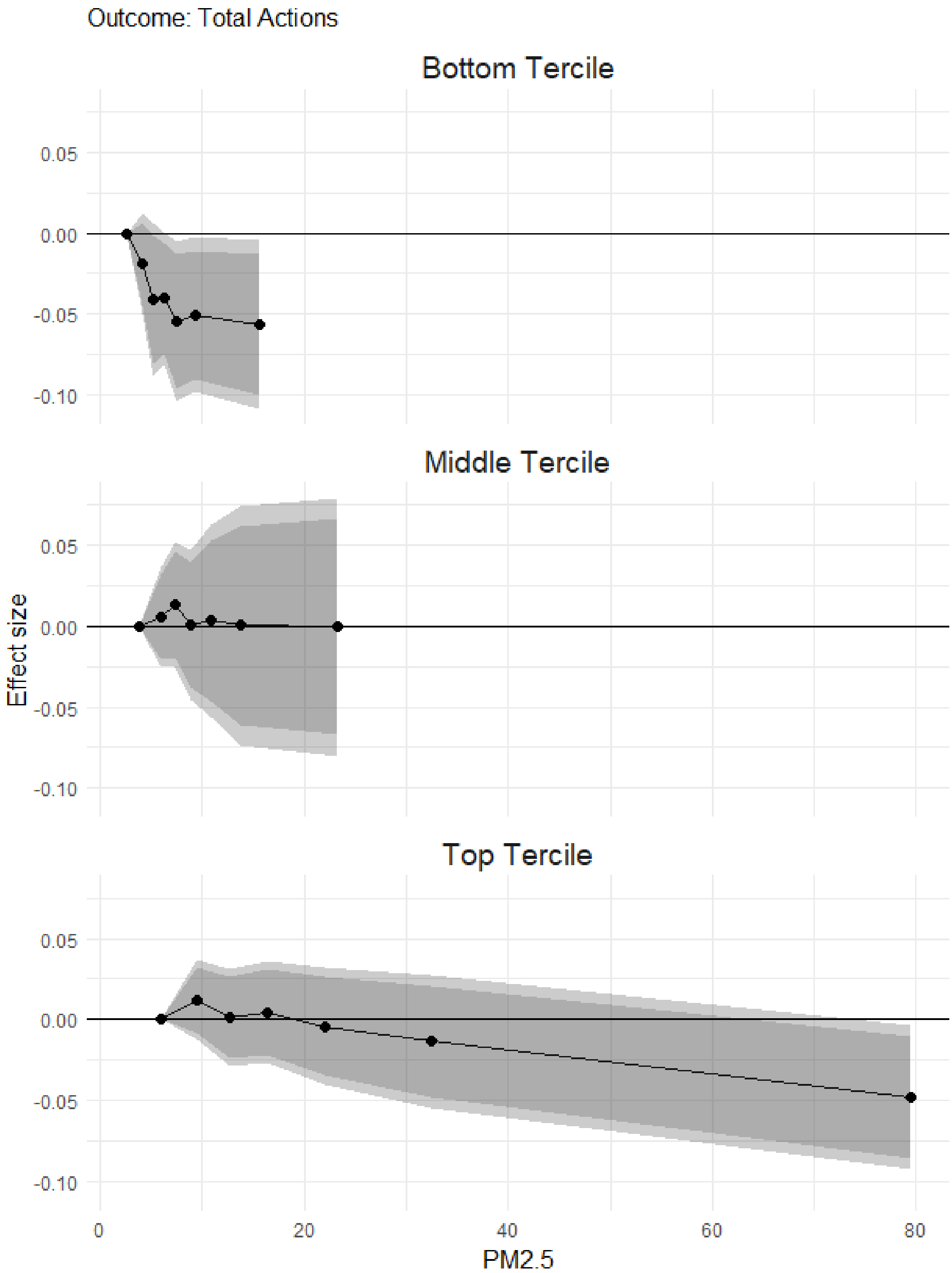


Figure 1.9.: Non-linear effects of $PM_{2.5}$ on Work Quantity across subsamples based on average $PM_{2.5}$

Note: The Figure depicts point estimates on different bins of $PM_{2.5}$ concentrations from OLS regressions of total actions on indicators for each bin for three distinct samples. Cities are assigned into subsamples based on average $PM_{2.5}$ concentration. Covariates: Weather and holiday controls as in Equation 1.1, region×date and city×month fixed effects. X-axis: Average $PM_{2.5}$ concentration in each bin in $\mu g/m^3$. Shaded areas indicate 95%- and 90%-confidence intervals.

ted variable bias. Moreover, individuals in high-pollution cities might engage more strongly in avoidance behavior and the use of protective devices such as air purifiers.[50] An important result, however, is that $PM_{2.5}$ exposure exerts adverse effects on productivity even below relatively strict current regulatory thresholds like those by the U.S. EPA. Given that the OLS estimates likely underestimate the true effects of $PM_{2.5}$ exposure, the results imply relevant economic benefits from complying with the stricter WHO standard for $PM_{2.5}$.

Effect Heterogeneity. Next, we analyze heterogeneity in the effect of fine particulate matter on work quantity by location characteristics in order to shed light on the distribution of air pollution damages and on potential mechanisms driving the adverse productivity effects.

We start by analyzing how the effect of an increase in $PM_{2.5}$ differs between places with low vs. high average pollution levels. To this end, we compute the average $PM_{2.5}$ concentration for each first-stage city group g and form two subsamples comprising city-groups with below and above median average $PM_{2.5}$ levels. We assign city-groups to the two subsamples instead of single cities to ensure that the IV approach does not capture impacts of local pollution transport such that we cleanly identify the causal effects of interest.

Average $PM_{2.5}$ concentration is 7.9 and 18.6 $\mu g/m^3$ in the two subsamples. Panel A of Table 1.8 presents the estimated effects on total actions and commits. We find larger point estimates in the low pollution sample. The impact on total actions, however, is not significant at conventional levels, likely due to the reduced sample size. This confirms the result from the OLS estimation of the dose-response function which also suggests stronger impacts at lower pollution levels. As mentioned above, more frequent use of air purifiers and other protective measures in high-pollution places might explain this result. Within the US, a similar pattern has been found by Bishop et al. (forthcoming) who analyze the impact of $PM_{2.5}$ on dementia.

Secondly, we investigate differences in effect magnitude between places with relatively high vs. low income levels. We collect data on GDP per capita in 2014, the first year of our sample period from the OECD, World Bank, and national statistical offices.[51] As before, we compute average values at the city-group level and assign groups into either the above or below median subsample. Average GDP per capita amounts to \$38,400 in the low-income sample, and \$71,000 in the high-income sample. Results are reported in Panel B of Table 1.8. We find that the point

[50] There is anecdotal evidence that big (tech) companies equip their offices in highly polluted places with air purifiers and filters, e.g., Microsoft, Google, SAP and Coca-Cola in Delhi or Nokia in Beijing.

[51] The main data source is the OECD's database on metropolitan areas, available at stats.oecd.org/Index.aspx?DataSetCode=CITIES. It provides GDP per capita for metropolitan areas, i.e., for some smaller cities in our sample we do not have city-specific data, but instead assign the value for the respective metro area. Small cities in Silicon Valley, e.g., Cupertino, Palo Alto, and Mountain View are assigned the GDP per capita reported for Greater San Francisco. Data for cities outside OECD countries is collected from national statistical agencies, the OECD regional statistics database, or the World Bank. Values are converted by the purchasing power parity conversion factor to adjust for differences in local price levels.

Table 1.8.: Effect Heterogeneity

	Actions (1)	*Commits* (2)	*Actions* (3)	*Commits* (4)
Panel A.	*Below Median $PM_{2.5}$*		*Above Median $PM_{2.5}$*	
$PM_{2.5}$	−0.0055	−0.0040**	−0.0027**	−0.0024**
	(0.0037)	(0.0016)	(0.0011)	(0.0009)
Observations	179,220	179,220	174,225	174,225
First Stage F-Stat.	107.4	107.4	102.4	102.4
Mean Dep. Var.	2.96	1.31	2.54	1.28
Mean $PM_{2.5}$	7.9	7.9	18.6	18.6
Panel B.	*Above Median GDP per capita*		*Below Median GDP per capita*	
$PM_{2.5}$	−0.0031	−0.0025**	−0.0030***	−0.0024**
	(0.0022)	(0.0011)	(0.0011)	(0.0010)
Observations	173,371	173,371	180,074	180,074
First Stage F-Stat.	88.9	88.9	113.4	113.4
Mean Dep. Var.	2.99	1.33	2.43	1.23
GDP	71,008	71,008	38,409	38,409
Mean $PM_{2.5}$	9.4	9.4	17.9	17.9

Note: Estimated coefficients reflect 2SLS estimates of the parameter β in Equation (1.1) for four distinct samples. The two samples used in Panel A are constructed by comparing average $PM_{2.5}$ concentration in each first stage city group g to the median value. The two samples used in Panel B are constructed by comparing average GDP per capita in 2014 in each first-stage city-group g to the median value. Data on per capita GDP is collected from the OECD, World Bank, and national statistical offices. The first stage specification is given in Equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation, and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week, and year-by-month fixed effects and the temperature controls can vary across world regions R. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01.

estimates for total actions and commits are of very similar magnitude in the two subsamples. In relative terms, the effects are marginally stronger in the high-income subsample.

Since GDP and air quality are systematically correlated, the two heterogeneity analyses cannot identify distinct drivers of effect magnitude.[52] The main takeaway message from the two heterogeneity analyses is that we find no evidence that the effects of air pollution on high-skilled worker productivity are more severe in disadvantaged locations. If anything, effects are stronger in places with better air quality. Health impacts of particulate pollution, on the other hand, are typically found to be larger in lower income and high-pollution locations (Colmer et al., 2021a; Hsiang et al., 2019). A potential explanation for why we do not find the same

[52] In Appendix Figure 1.B.4, we present the distribution of average $PM_{2.5}$ concentration in the rich vs. poor city groups and the distribution of GDP per capita in the clean vs. polluted city groups. Lower-income locations on average have higher pollution levels.

pattern could be that we consider only individuals in a high-paying occupation.

Next, we investigate whether effect magnitude differs between places with low vs. high awareness of air pollution as an important issue. We use data from the Pew Research Center International Science Survey which was conducted in early 2020 across 20 countries with several thousand interviewees per country. Survey participants were asked whether they believe that air pollution is a big, a moderate, a small, or no problem at all in their country. As a country-wide measure of awareness, we compute the share of respondents stating that air pollution is a big problem. Appendix Figure 1.B.5 shows the distribution of this variable across the 14 countries that are included in both our data and the survey and how it varies with average $PM_{2.5}$ concentration. 152 cities in our main sample are covered by the survey data. We split these cities into three groups with low, intermediate, and high awareness.[53] Table 1.9 presents 2SLS estimates for the effect of $PM_{2.5}$ concentration on total actions for the total sample covered by the survey data, and the three subsamples. Effects are negative and significant across all samples, and importantly there is no clear gradient in awareness. In fact, the point estimate is identical in the high and the low awareness samples. This suggests that the reduction in output is not driven by avoidance behavior, e.g., working from home on high pollution days which reduces productivity. In this case, we would expect to see larger effects in the high awareness sample.

Lastly, we investigate effect heterogeneity based on the quality of the local building stock. Effective exposure to particulate matter is likely lower for individuals inside modern buildings with low penetration rates than for individuals in older, lower-quality buildings given the same outdoor concentration. We use data on the construction period of residential dwellings as a proxy for building stock quality. We collect data on building stock age from different national statistical offices, covering 164 out of the 193 cities in our main sample.[54] For each first-stage cluster, we compute the average share of dwellings built before 1970, i.e., the share of relatively old buildings which likely have high indoor penetration rates. As before, we then assign groups into subsamples based on whether the group level share of old buildings is above or below the median value. Panel B of Table 1.9 presents regression results for the full sample covered by the building stock data, as well as the two subsamples with above and below median share of old buildings. We find that the negative effect of $PM_{2.5}$ on total actions is driven by the sample with a relatively high share of old dwellings, whereas the point estimate

[53] US cities form the intermediate awareness sample. Among US respondents, 63.1% believe that air pollution is a big problem. All countries where a larger (smaller) share of respondents holds this view, are assigned to the high (low) awareness subsample.

[54] The data is collected from the American Community Survey for metropolitan areas in the US, the EU Building Stock Observatory for EU member states (country-level), the Federal Statistical Office of Switzerland (canton-level), the Statistics Bureau of Japan (prefecture-level), Statistics Canada (province-level), and Statistics Norway (municipality-level).

Table 1.9.: Heterogeneity: Awareness and Building Stock Age

	Actions			
Panel A. Awareness				
	Total	*High Awareness*	*Intermediate Awareness*	*Low Awareness*
$PM_{2.5}$	-0.0043^{***}	-0.0035^{**}	-0.0083^{*}	-0.0035^{*}
	(0.0012)	(0.0014)	(0.0042)	(0.0020)
Observations	280,297	80,963	120,521	78,813
First Stage F-Stat.	98	90	67	115
Share AP is Big Problem	67.4%	78.1%	63.1%	51.6%
Mean $PM_{2.5}$	11.0	16.2	8.7	9.5
Mean Dep. Var.	2.9	2.6	2.9	3.1
Panel B. Building Stock Age				
	Total	*Above Median Old Building Share*	*Below Median Old Building Share*	
$PM_{2.5}$	-0.0039^{**}	-0.0044^{***}	-0.0025	
	(0.0015)	(0.0016)	(0.0031)	
Observations	300,844	167,307	133,537	
First Stage F-Stat.	102	124	93	
Share modern buildings	28%	19%	38%	
Share old buildings	44%	55%	32%	
Mean $PM_{2.5}$	10.4	9.7	10.9	
Mean Dep. Var.	2.92	2.98	2.90	

Note: Estimated coefficients reflect 2SLS estimates of the parameter β in Equation (1.1) where the outcome variable is the number of completed actions. Each Column is estimated on a different sample. In Panel A, the sample used in Column (1) includes all 153 cities covered by the Pew Research Center International Science Survey. Results in Columns (2) to (4) are estimated on subsamples formed based on country-level awareness of air pollution, measured by the share of respondents stating that air pollution is a big problem in the Pew Survey. In Panel B, the sample used in Column (1) includes all 169 cities covered by data on building stock age. Results in Columns (2) to (3) are estimated on subsamples formed based on the city-group level share of dwellings built before 1970, which are defined as old buildings. Modern buildings are those build after 1990. The first stage specification is given in Equation (1.2). Covariates are as described in Table 1.8. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01.

is not statistically significant and less than 60% as large in the subsample with relatively few old buildings. The fact that effects are larger in places where effective exposure is likely higher suggests that the main results are driven by physiological effects of air pollution, rather than by behavioral changes or avoidance behavior.

Monthly Level. Next, we quantify the effect of $PM_{2.5}$ on output at a more aggregate time period, by estimating a model at the developer×month level. This is motivated by the findings that pollution exposure has not only contemporaneous, but also some lagged effects on output, and that developers partially compensate for the adverse productivity shocks by working more on weekends. We analyze effects on the same output quantity measures used before (actions, commits, and comments). In addition, we consider effects on the growth rate of the number of developers' followers. We view this as a proxy for the quantity, quality, and relevance of a developer's work on GitHub because all these dimensions likely affect the decision of other GitHub users about whether to follow the developer or not.

To explore effects at the monthly level, we need to adapt the IV strategy. We use three variables measuring the share of days in month m with wind direction falling into a specific 90° bin, each interacted with indicators for the first-stage city-groups g, as instruments.

Table 1.10 presents the results. An increase in monthly $PM_{2.5}$ concentration by 1 μg/m³ reduces the number of actions performed in that month by 0.17 or 0.21% of the mean. The implied effect of an increase in $PM_{2.5}$ by one μg/m³ on a single day is .0057, and thus slightly larger than the effect found in the analysis at the daily level (.0032). Again, this effect is mostly driven by a reduction in commits, which fall by 0.31% of the sample mean, while the reduction in comments is small and insignificant. $PM_{2.5}$ also negatively affects the growth rate of the number of followers. The estimate implies a decrease of 1.5% relative to the mean rate. In sum, exposure to air pollution has negative impacts on developers' output also over a more aggregate time period. It even slows down the process of gaining reputation in the tech community, which might have adverse long-run consequences for developers' career paths.[55]

Robustness Checks. In Appendix Tables 1.A.12 to 1.A.13 we show that our main results are not sensitive to specific choices on how we set up the first and second stage models. We find evidence for a reduction in output quantity—driven by fewer commits, but no or very small changes in comments—and a switch towards easier issues and PRs in response to $PM_{2.5}$ across specifications.

First, we examine robustness to the specification of the wind direction instruments. Instead of $\sin\left(\theta_{c,d}\right)$ and $\sin\left(\theta_{c,d}/2\right)$, we use three indicator variables for average daily wind direction falling into a specific 90° bin (south-west, south-east and north-east, with north-west as omitted category), following Deryugina et al. (2019). The results are reported in Panels A and B of Table 1.A.12. In Panels C and D, we report results from a specification where we used a k-means clustering algorithm, instead of hierarchical clustering, to form the city-groups g across

[55]Inspecting GitHub pages of potential employees is a common practice in hiring decisions in the tech sector as described in several tech blogs, see e.g., https://techbeacon.com/app-dev-testing/what-do-job-seeking-developers-need-their-github or https://blog.boot.dev/jobs/build-github-profile/.

Table 1.10.: Analysis at the Monthly Level

	Actions (1)	*Commits* (2)	*Comments* (3)	*Growth Rate(Followers)* (4)
$PM_{2.5}$ (monthly)	-0.173^{**}	-0.125^{***}	-0.033	-0.00011^{***}
	(0.076)	(0.035)	(0.045)	(0.00003)
	[0.024]	[0.0005]	[0.474]	[0.0004]
F-Statistics	644	644	644	636
Observations	469,373	469,373	469,373	453,443
Mean Dep. Var.	84.3	39.3	28.3	.0072

Note: The table presents IV estimates of the effect of monthly $PM_{2.5}$ concentration on the outcomes described at the top of the Table. The excluded instruments are variables measuring the share of days in the month on which wind direction was blowing from one of three 90° angles, interacted with indicators for first-stage city groups g. Regressions control for developer and region-by-year-by-month fixed effects, third-order polynomials in average monthly temperature, precipitation, relative humidity, and wind speed, the number of holidays and days with heavy wildfire smoke at the city×month level. Temperature controls and effects of holidays are allowed to vary across regions R. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. *p<0.1; **p<0.05; ***p<0.01

which the effects of wind direction are allowed to differ in the first stage. In both cases, results on work quantity and task choice are very similar to the baseline results.

Second, we test robustness to the functional form chosen in the second stage model. Table 1.A.13 shows the estimated effects of $PM_{2.5}$ when work output is measured by the inverse hyperbolic sine transformation of total actions, commits, and comments, respectively (Panels A and B). Again, the direction and statistical significance of the baseline results persist, but this specification implies somewhat smaller effect magnitudes. Panel C displays results when $PM_{2.5}$ in logs is used as regressor. This yields a high F-Statistic and the same pattern for second-stage effects on work quantity and task complexity as the baseline model.

In Table 1.A.14 we show that the statistical significance of our results persists if we cluster standard errors at the level of the city-groups g across which the effects of wind direction are allowed to differ in the first stage, instead of the city level.

Lastly, we demonstrate that the results are overall robust to changes in the included fixed effects and weather conditions. Tables 1.A.15 to 1.A.18 show that across specifications with different fixed effects absorbing common time shocks at different geographic and temporal levels, and across specifications with more and less detailed weather controls, our main results hold.

Extended Sample. In Table 1.A.19 we show that our core results also hold in the extended sample including all 220 cities in our dataset and with instruments based on temperature inversions instead of wind direction. Specifically, we estimate the model in Equation 1.1, but use a variable measuring inversion strength (as specified in Section 1.3) interacted with indicator variables for geographic regions r as instruments. We allow effects of inversions to vary geographically because the strength of the first stage effect varies based on baseline emissions (Krebs and Luechinger, 2021). We form 15 regions r to make sure that each region comprises multiple cities and forms a homogenous geographic area.[56] In Appendix Table 1.A.19, we present results for the outcomes measuring output quantity (Columns 1 to 3) and the outcomes measuring whether developers switch to less complex tasks (Columns 4 to 6). Apart from the effect on the share of issue events referring to an easy issue, which is small and insignificant, results replicate the patterns we found in our main analysis and are of comparable magnitude.

1.7. Conclusion

How do environmental conditions, like fluctuations in air pollution, affect workers in jobs that form the backbone of the modern knowledge economy? These jobs are focused on interpersonal and analytic tasks, often require strong social and digital skills, and are organized in a way that gives workers flexibility in schedules and task choices. As digitization and automation continue to change the world of work, these job characteristics are expected to become even more widespread.

In this paper, we use detailed data from GitHub to study how particulate matter affects daily output and work patterns in a global sample of software developers—a high-skilled occupation that can be considered as representative of the jobs of interest.

We provide evidence that pollution exposure reduces developer output. On a day with unusually high pollution ($PM_{2.5}$ concentration above the city-specific 75th percentile) the total number of actions conducted by software developers falls by 4% relative to days with better air quality. This effect is mostly driven by a reduction in *individual* coding activity, while the level of *collaborative* activity is unaffected. Our estimates are at the lower end of air pollution effects found in other, less flexible and less collaborative occupations studied in previous research. Moreover, we find no evidence of a deterioration in output quality. Due to the high value generated by software developers, the implied monetary loss is nevertheless economically relevant and comparable to findings for workers in manual occupations. Our estimates imply that on a day with unusually high pollution, output value falls by $11 per developer.

Our second key result is that software developers exploit the flexibility of their work setting to adapt to increases in air pollution. In particular, we find that they choose to work on less

[56] These groups are the four US census regions, Canada, China, India, South East Asia, Japan and South Korea, Australia, New Zealand, Western, Eastern, Southern, and Northern Europe.

complex tasks when $PM_{2.5}$ increases. Among developers who respond with a stronger shift towards easier tasks, effects on output quantity are alleviated. In addition, developers reallocate work activity from high-pollution, low-productivity workdays to low-pollution, high-productivity weekends. One additional day with unusually high $PM_{2.5}$ concentration in the first half of the week causes an increase in weekend work by 2.1%. These forms of adaptation likely explain why the effects on output quantity and quality in our setting are small relative to previous studies. At the same time, they suggest an additional welfare cost of air pollution in this setting not captured by changes in output due to forgone leisure time on the weekend and potential negative impacts on work-life balance.

While we use data on a sample of software developers who use GitHub as part of their professional work, we believe that the findings are externally valid to workers in many other high-paying occupations which offer flexible schedules and discretion in task choice, and require similar skills, e.g., problem-solving skills, attention to detail, programming, and teamwork. This applies to many high-skill workers, including business analysts or researchers. Furthermore, the fact that our data comprises developers across more than 30 countries suggests that the effects we identify and quantify in this study are not specific to a certain firm or country context but apply more generally. Based on this, we can derive estimates of the monetary benefits in terms of productivity gains among knowledge workers from reducing $PM_{2.5}$ concentration permanently by one unit. Extrapolating to all U.S. workers in the occupation group "Computer and Mathematical Workers" and to all ICT professionals in the EU suggests annual benefits of \$580m (US) and \$980m (EU), respectively.[57]

Hence, our findings have important policy implications. When deciding about limit values on air pollutants, regulators should take the growing evidence on the economic benefits of pollution reductions in the form of productivity gains into account. Importantly, we find that adverse effects of $PM_{2.5}$ on output are large at concentrations below the regulatory standards in force in the European Union and the US. Hence, even in areas with relatively good air quality, further improvements will likely generate additional benefits. While we find slightly smaller marginal effects in high pollution locations, the fact that $PM_{2.5}$ concentration is often an order of magnitude larger in developing countries like India and Bangladesh compared to the US might be an important barrier to growth for the software industries in these countries.

Our findings on how software developers adjust work patterns also have interesting impli-

[57] Our 2SLS estimates for the effect of $PM_{2.5}$ concentration on commits and PRs, paired with the estimates of the monetary value of these outcomes, imply that a one unit decrease in $PM_{2.5}$ increases daily output value by \$0.344 per developer. Employment in "Computer and Mathematical Workers" in the US in 2021 was 4,654,750 (Bureau of Labor Statistics, 2022). Per year the total estimated gain in output value in this group is thus \$0.344/worker and day $\times$ 4,654,750 workers $\times$ 365 days = \$580m, if we assume that the effect of a permanent reduction in $PM_{2.5}$ is given by the sum of the daily effects. We compute the value for the EU analogously, based on an employment figure of 7,843,000 ICT professionals in 2020 (Cedefop, 2022).

cations for the organization of work within firms: Highlighting the difficulty of certain tasks, as done by the use of issue labels on GitHub, and granting flexibility in working hours, might help workers to better adapt to idiosyncratic productivity shocks and mitigate the total impact on team or firm performance.

Appendix to Chapter 1

1.A. Additional Tables

Table 1.A.1.: Characteristics of High-Skill Occupations and Software Development

	Freedom to Make Decisions		Structured versus Unstructured Work		Work With Work Group or Team	
	All high-skill occupations	Software developers	All high-skill occupations	Software develop.	All high-skill occupations	Software develop.
1	0.4	0	0.6	0	1.79	0
2	2.6	3.1	2.3	2.4	4.4	5.9
3	10.5	29.1	11.3	28.1	11.0	2.7
4	35.6	38.2	39.8	45.0	30.5	9.2
5	50.9	29.6	46.0	24.6	52.4	82.3

Note: Based on data from O*NET Database Version 25.0. Work Contexts Table. All high-skill occupations refers to occupations in Job Zones 4 and 5. Software developers refers to occupation 15-1132.00 ("Software Developers, Applications"). Categories: 1 = not important at all/no freedom; 2 = Fairly important/very little freedom; 3 = Important/Limited freedom; 4= Very Important/Some freedom; 5 = Extremely important / A lot of freedom

Table 1.A.2.: Labels Indicating *Easy* Issues

good first issues	good first bug	good-first
documentation	polish	cleanup
simple	easy	small
trivial	minor	**help wanted**
junior job	newcomer	starter
beginner	newbie	novice
low hanging	low-hanging	

Note: If a label contains any of these terms, the issue is classified as "easy". Bolt text indicates GitHub default labels.

Table 1.A.3.: Description of Outcome Variables

Domain	Concept	Variable	Details
Output quantity	Total output quantity	Actions	Sum of number of commits, comments on issues, PRs and commits, PRs opened, PRs closed, issues opened, closed and reopened
	Coding activity	Commits	Number of commits
	Interactive activity	Comments	Sum of number of comments written on issues, PRs and commits
Output Quality	PR Success rate	Share PRs merged	PRs opened that got merged/all PRs opened
	Deficient commits	Share commits reverted	Commits that got reverted/all commits
Task choice	Easy tasks among issue events	Share easy issue events issues	(#easy issues opened + #easy issues closed + #comments written on easy issues)/(#issues opened + #issues closed + #comments written on issues)
	Average PR complexity	Lines added per PR	Average number of lines of code added in PRs opened, closed and commented on
		Files changed per PR	Average number of code files changed in PRs opened, closed and commented on
Working hours	Evening activity	Time last action	Minute of final action of the day
		Share of actions after 6 pm	Actions made after 6 pm/Total actions

Note: The Table displays information on the outcome variables we use, how they are constructed, and what they measure.

Table 1.A.4.: Sources of Air Quality Data

Geographic Area	**Data Source**
United States	U.S. Environmental Protection Agency (EPA)
Canada	Canadian National Air Pollution Surveillance (NAPS) Program
Mexico City	Gobierno de la Ciudad de México
Europe	European Environment Agency (EEA)
Russia, Ukraine, Belarus, Turkey, Israel	Copernicus Atmosphere Monitoring Service (CAMS)
China	National Environmental Monitoring Centre
Mumbai Hyderabad Chennai New Delhi Dhaka	US Embassies (AirNow.gov)
Bengaluru	Central Pollution Control Board (CPCB)
Japan	National Institute for Environmental Studies
Hong Kong	Hong Kong Environmental Protection Department
Singapore	National Environment Agency
South Korea	Air Korea
Taiwan	Environmental Protection Administration
Australia	New South Wales Department of Planning and Environment Victorian Government open data portal Queensland Government open data portal South Australian Government Data Directory
New Zealand	Stats NZ Tatauranga Aotearoa

Note: Data sources for data on $PM_{2.5}$. Airbase, the EEA's database on air pollution, contains monitor data for 33 countries, including all EU members, as well as further EEA member and cooperating countries, e.g., Switzerland, Norway and Serbia.

Table 1.A.5.: Distribution of developer-by-date observations across geographic regions R

Region R	Observations	Share
Oceania	273,246	1.9
Northern America	7,244,272	50.6
Northern Europe	1,809,844	12.6
Western Europe	2,312,377	16.1
Southern Europe	333,597	2.3
Eastern Europe	725,664	5.1
Asia	1,628,930	11.4

Note: The table shows the distribution of observations in the developer × date panel described in section 1.3.2 across geographic regions R.

Table 1.A.6.: Effect of $PM_{2.5}$ on Quantity of Issue and Pull Request Actions

	PRs closed (1)	*PRs opened* (2)	*Issues closed* (3)	*Issues opened* (4)
Panel A.				
$PM_{2.5}$	-0.00011	-0.00018^{**}	0.00004	0.00009
	(0.00012)	(0.00008)	(0.00009)	(0.00006)
First Stage F-Stat.	102.1	102.1	102.1	102.1
% change in Y	-0.1	-1.2	0.03	0.1
Panel B.				
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.0049^{*}	-0.0064^{***}	0.0005	-0.0023
	(0.0029)	(0.0023)	(0.0029)	(0.0024)
First Stage F-Stat.	80.5	80.5	80.5	80.5
% change in Y	-2.9	-4.2	0.4	2.2
Mean Dep. Var.	0.17	0.15	0.12	0.11
Observations	353,445	353,445	353,445	353,445

Note: The table presents IV estimates of the parameter β in equation (1.1). In Panel A, the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Panel B, a binary variable is used instead, which takes a value of one if city×day $PM_{2.5}$ concentration exceeds the city-specific 75th percentile. The first stage specification is given in equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and the temperature controls can vary across world regions R. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1.A.7.: OLS Results for Work Quantity (main sample)

	Actions (1)	*Commits* (2)	*Comments* (3)	*Actions* (4)	*Commits* (5)	*Comments* (6)
Panel A.						
$PM_{2.5}$	−.00024	−.0002	−.00001	−.0006*	−.0003	−.0002*
	(.0003)	(.0002)	(.0001)	(.0003)	(.0002)	(.0001)
Panel B.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	−.0158*	−.0096**	−.0026	−.0264**	−.0131**	−.0054
	(.0084)	(.0041)	(.0036)	(.0115)	(.0051)	(.0049)
Observations	353,445	353,445	353,445	353,445	353,445	353,445
City FE	✓	✓	✓			
Region×Day-of-Week FE	✓	✓	✓			
Region×Year-Month FE	✓	✓	✓			
Region×Date FE				✓	✓	✓
City×Month FE				✓	✓	✓

Note: The table presents OLS estimates of the parameter β in equation (1.1), where the dependent variables are displayed in the upper part of the table. Parameters are estimated on the main sample including 193 cities. The regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$ in Panel A, and an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile in Panel B. Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation and relative humidity, indicators for heavy wildfire smoke and holidays. The temperature controls can vary across world regions R. Included fixed effects are displayed in the bottom part of the table. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.8.: OLS Results for Work Quantity (extended sample)

	Actions (1)	*Commits* (2)	*Comments* (3)	*Actions* (4)	*Commits* (5)	*Comments* (6)
Panel A.						
$PM_{2.5}$	$-.0004^{***}$	$-.0002^{**}$	$-.0002^{**}$	$-.0004^{**}$	$-.0002$	$-.0001^{***}$
	(.0001)	(.0001)	(.0001)	(.0002)	(.0001)	(.00004)
Panel B.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	$-.0160^{**}$	$-.0087^{**}$	$-.0036$	$-.0234^{**}$	$-.0119^{**}$	$-.0049$
	(.0078)	(.0039)	(.0032)	(0.0102)	(.0046)	(.0042)
Observations	398,687	398,687	398,687	398,687	398,687	398,687
City FE	✓	✓	✓			
Region×Day-of-Week FE	✓	✓	✓			
Region×Year-Month FE	✓	✓	✓			
Region×Date FE				✓	✓	✓
City×Month FE				✓	✓	✓

Note: The table presents OLS estimates of the parameter β in equation (1.1), where the dependent variables are displayed in the upper part of the table. Parameters are estimated on the extended sample including 220 cities. The regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$ in Panel A, and an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile in Panel B. Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation and relative humidity, indicators for heavy wildfire smoke and holidays. The temperature controls can vary across world regions *R*. Included fixed effects are displayed in the bottom part of the table. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.9.: Reduced Form

	Actions (1)	*Commits* (2)	*Comments* (3)
High-Pollution Wind Direction	-0.0192^{**}	-0.0111^{***}	-0.0036
	(0.0087)	(0.0040)	(0.0038)
	[0.029]	[0.006]	[0.350]
Observations	367,472	367,472	367,472
First Stage Effect on $PM_{2.5}$		3.683^{***} (0.456)	

Note: The Table displays OLS estimates of the outcomes displayed in the upper part of the table on an indicator variable for wind blowing towards a city from the direction (60° angle) that has the largest positive effect on local $PM_{2.5}$ concentration. Standard errors clustered at the city level are reported in parentheses. P-values are presented in squared brackets. All regressions include covariates as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.10.: Effect of $PM_{2.5}$ on PRs opened and closed with GHArchive and GHTorrent data

	PRs opened (GHA) (1)	*PRs closed (GHA)* (2)	*PRs opened (GHT)* (3)	*PRs closed (GHT)* (4)
Panel A.				
$PM_{2.5}$	-0.00029^{**}	-0.00007	-0.00016^{*}	-0.00011
	(0.00014)	(0.00013)	(0.00009)	(0.00013)
	[0.036]	[0.581]	[0.067]	[0.378]
First Stage F-Stat.	89	89	89	89
Panel B.				
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.00668^{*}	-0.00458	-0.00701^{***}	-0.00586^{**}
	(0.00377)	(0.00297)	(0.00239)	(0.00297)
	[0.078]	[0.124]	[0.004]	[0.050]
First Stage F-Stat.	69	69	69	69
Observations	298,566	298,566	298,566	298,566

Note: The table presents IV estimates of the parameter β in equation (1.1), where the outcome is the number of pull requests (PRs) opened or closed, respectively. This is measured based on GHArchive data in columns (1) to (2) and GHTorrent data in column (3) to (4). The regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$ in Panel A, and an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile in Panel B. The first stage specification is given in equation (1.2). Covariates include eight bins for mean daily temperature, third-order polynomials in wind speed, precipitation and relative humidity, indicators for heavy wildfire smoke and holidays, as well as city, day-of-week, and year-by-month fixed effects. Day-of-week and year-by-month fixed effects and the temperature controls can vary across world regions R. The sample period is 2015 to May 2019. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.11.: Placebo Test: Effect of $PM_{2.5}$ Friday to Sunday on Work Activity Wednesday to Thursday

	Actions		*Commits*		*Comments*	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A.						
$PM_{2.5}$	0.0065	0.0055	0.0045	0.0028	0.0006	0.0017
	(0.0092)	(0.0086)	(0.0039)	(0.0041)	(0.0035)	(0.0034)
Panel B.						
High $PM_{2.5}$ Days	0.0401	−0.0082	0.0293	−0.0020	0.0034	−0.0007
	(0.0888)	(0.0795)	(0.0364)	(0.0352)	(0.0346)	(0.0316)
Observations	1,997,123	1,321,642	1,997,123	1,321,642	1,997,123	1,321,642
Weeks	all	only low PM weekends	all	only low PM weekends	all	only low PM weekends

Note: The table presents IV estimates of the parameter β in a placebo version of equation 1.4. Outcomes are the sum of all actions, commits and comments made between Wednesday and Thursday, the placebo weekend. In Panel A, the regressor of interest is average $PM_{2.5}$ concentration between Friday and Sunday of the week before. In Panel B, the count of of days on which the city×day $PM_{2.5}$ concentration exceeds the city-specific 75th percentile during this period is used instead. The first stage specification is given in equation 1.2. Regressions control for developer and region-by-year-by-quarter fixed effects, the number of public holidays during the workweek, and the leads of the instrumental variables for both the placebo weekend and the period from Monday to Tuesday. Further covariates are the number of days with heavy wildfire smoke, and third-order polynomials in average wind speed, precipitation, and relative humidity during both the placebo weekend and the exposure period between Friday and Sunday. Temperature controls are included in the form of eight bin variables for the placebo exposure period, and in the form of a third order polynomial for the placebo weekend, and are allowed to vary across regions R. Standard errors clustered at the city level are reported in parentheses. P-values are reported in brackets. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1.A.12.: Robustness: First Stage Specification

	Actions (1)	*Commits* (2)	*Comments* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A.						
$PM_{2.5}$	-0.0018^{*}	-0.0020^{***}	0.0001	0.0001^{*}	-0.0019^{**}	-0.0012^{***}
	(0.0010)	(0.0007)	(0.0005)	(0.0001)	(0.0009)	(0.0004)
First Stage F-Stat.	62	62	62	52	38	38
Panel B.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.0629^{**}	-0.0661^{***}	0.0059	0.0031^{**}	-0.0469^{**}	-0.0285^{***}
	(0.0290)	(0.0128)	(0.0182)	(0.0016)	(0.0201)	(0.0100)
First Stage F-Stat.	49	49	49	41	41	28
IV-Specification	Three wind direction bins					
Clustering	Hierarchical Clustering Algorithm					
Observations	353,445	353,445	353,445	250,376	164,883	164,883
Panel C.						
$PM_{2.5}$	-0.0030^{***}	-0.0024^{***}	-0.0005	0.0001^{**}	-0.0012	-0.0008^{**}
	(0.0011)	(0.0006)	(0.0006)	(0.0001)	(0.0008)	(0.0004)
First Stage F-Stat.	102	102	102	86	62	62
Panel D.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.1023^{***}	-0.0739^{***}	-0.0160	0.0029^{**}	-0.0347^{*}	-0.0201^{**}
	(0.0268)	(0.0144)	(0.0178)	(0.0014)	(0.0177)	(0.0095)
First Stage F-Stat.	81	81	81	66	46	46
IV-Specification	$\sin(\theta), \sin(\frac{\theta}{2})$					
Clustering	K-means Clustering Algorithm					
Observations	353,445	353,445	353,445	250,376	164,883	164,883

Note: The table presents IV estimates of the parameter β in Equation (1.1). In Panels A and C, the regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$. In Panel B and D, an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. Relative to specifications underlying results in Table 1.2, the first stage model is changed: In Panels A and B, instruments are three indicators for wind direction falling in specific bins, each covering 90° of the wind rose. In Panels C and D, the first stage specification is as in Equation (1.2), but we form city-groups g using k-means clustering instead of hierarchical clustering. Covariates as described in Table 1.2. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.13.: Robustness: Second Stage Specification

	Actions (1)	*Commits* (2)	*Comments* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A. Inv. Hyperbolic Sine Transformation						
$\text{PM}_{2.5}$	-0.0005^{**}	-0.0005^{**}	-0.0001			
	(0.0002)	(0.0002)	(0.0001)			
First Stage F-Stat.	102	102	102			
Panel B. Inv. Hyperbolic Sine Transformation						
$\mathbb{1}\{\text{PM}_{2.5} > Q_{0.75}\}$	-0.0206^{***}	-0.0159^{***}	-0.0082^{**}			
	(0.0060)	(0.0043)	(0.0039)			
First Stage F-Stat.	80	80	80			
Panel C. log(PM)						
$log(\text{PM}_{2.5})$	-0.0656^{***}	-0.0444^{***}	-0.0141	0.0018^{**}	-0.0209^{**}	-0.0143^{***}
	(0.0163)	(0.0087)	(0.0090)	(0.0009)	(0.0096)	(0.0052)
First Stage F-Stat.	197	197	197	162	117	117
Observations	353,445	353,445	353,445	250,376	164,883	164,883

Note: The table presents IV estimates of the parameter β in Equation (1.1). In Panel A, the regressor of interest is $\text{PM}_{2.5}$ concentration measured in $\mu g/m^3$. In Panel B, an indicator for $\text{PM}_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. In Panel C, the regressor is the logarithm of $\text{PM}_{2.5}$ concentration. Inverse hyperbolic sine transformations are applied to outcomes in Panels A and B. The first stage specification is given in Equation (1.2). Covariates as described in Table 1.2. Regressions are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.14.: Robustness: Clustering of Standard Errors

	Actions (1)	*Commits* (2)	*Comments* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A.						
$PM_{2.5}$	-0.0032^{***}	-0.0026^{***}	-0.0005	0.0001^{**}	-0.0013	-0.0010^{**}
	(0.0012)	(0.0007)	(0.0006)	(0.0001)	(0.0008)	(0.0004)
	[0.009]	[0.0003]	[0.421]	[0.031]	[0.110]	[0.031]
Panel B.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.1104^{***}	-0.0801^{***}	-0.0169	0.0032^{**}	-0.0403^{**}	-0.0244^{**}
	(0.0301)	(0.0154)	(0.0189)	(0.0013)	(0.0171)	(0.0092)
	[0.001]	[0.000004]	[0.379]	[0.017]	[0.023]	[0.011]
Observations	353,445	353,445	353,445	250,376	164,883	164,883

Note: The Table presents IV estimates of the parameter β in Equation (1.1). In Panels A, the regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$. In Panel B, an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. The first stage specification is given in Equation (1.2). Standard errors clustered at the level of *city-groups g* across which the effect of instruments in the first stage are allowed to differ are reported in parentheses. P-values are presented in squared brackets. All regressions include covariates as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.15.: Robustness to Changes in Weather controls (Output Quantity)

	Actions (1)	*Commits* (2)	*Comments* (3)	*Actions* (4)	*Commits* (5)	*Comments* (6)
Panel A.						
	-0.0041^{***}	-0.0027^{***}	-0.0009^{**}	-0.1156^{***}	-0.0749^{***}	-0.0261^{**}
	(0.0010)	(0.0007)	(0.0004)	(0.0269)	(0.0150)	(0.0115)
First Stage F-Stat.	147	147	147	110	110	110
Weather Controls	none					
Panel B.						
	-0.0029^{***}	-0.0022^{***}	-0.0005	-0.1162^{***}	-0.0767^{***}	-0.0243
	(0.0010)	(0.0007)	(0.0005)	(0.0316)	(0.0159)	(0.0153)
First Stage F-Stat.	108	108	108	83	83	83
Weather Controls	Quadratic functions of precipitation, wind speed, rel. humidity and cubic, continent-specific function of mean temperature					
Panel C.						
	-0.0018	-0.0021^{**}	0.0001	-0.0539	-0.0601^{***}	0.0069
	(0.0012)	(0.0009)	(0.0006)	(0.0345)	(0.0190)	(0.0190)
First Stage F-Stat.	87	87	87	67	67	67
Weather Controls	Continent specific bins for precipitation, wind speed, rel. humidity, minimum and maximum temperature					
Observations	353,445	353,445	353,445	353,445	353,445	353,445

Note: The table presents IV estimates of the parameter β in Equation (1.1). Dependent variables are denoted at the top of the table. In Columns(1) to (3), the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Columns (4) to (6), an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. Relative to specifications underlying results in Table 1.2, we change the included covariates to control for weather conditions. We state the included variables at the bottom of each Panel. The first stage specification is given in Equation (1.2). All regressions include fixed effects as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.16.: Robustness to Changes in Fixed Effects (Output Quantity)

	Actions (1)	*Commits* (2)	*Comments* (3)	*Actions* (4)	*Commits* (5)	*Comments* (6)
Panel A.						
	-0.0044^{***}	-0.0031^{***}	-0.0009	-0.1438^{***}	-0.0954^{***}	-0.0288
	(0.0013)	(0.0009)	(0.0006)	(0.0337)	(0.0169)	(0.0189)
First Stage F-Stat.		82			66	
Fixed Effects	city, region × day-of-week, region × week					
Panel B.						
	-0.0045^{***}	-0.0034^{***}	-0.0008	-0.1323^{***}	-0.0995^{***}	-0.0170
	(0.0014)	(0.0009)	(0.0007)	(0.0399)	(0.0207)	(0.0215)
First Stage F-Stat.		84			62	
Fixed Effects	city, region × date					
Panel C.						
	-0.0032^{**}	-0.0021^{***}	-0.0007	-0.1018^{***}	-0.0603^{***}	-0.0279^{**}
	(0.0013)	(0.0008)	(0.0006)	(0.0257)	(0.0142)	(0.0127)
First Stage F-Stat.		101			76	
Fixed Effects	city × month, region × day-of-week, region × year × month					
Panel D.						
	-0.0048^{**}	-0.0028^{**}	-0.0013^{*}	-0.1311^{***}	-0.0711^{***}	-0.0403^{***}
	(0.0019)	(0.0011)	(0.0008)	(0.0319)	(0.0181)	(0.0139)
First Stage F-Stat.		78			61	
Fixed Effects	city × month, region × day-of-week, region × week					
	-0.0023^{*}	-0.0012^{*}	-0.0008	-0.0762^{***}	-0.0369^{***}	-0.0283^{*}
	(0.0013)	(0.0006)	(0.0007)	(0.0289)	(0.0138)	(0.0160)
First Stage F-Stat.		114			86	
Fixed Effects	region × day-of-week, city × year, city × month					
Observations	353,445	353,445	353,445	353,445	353,445	353,445

Note: The table presents IV estimates of the parameter β in Equation (1.1). Dependent variables are denoted at the top of the table. In Columns(1) to (3), the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Columns (4) to (6), an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. Relative to specifications underlying results in Table 1.2, we change the included fixed effects. We state the included fixed effects at the bottom of each Panel. The first stage specification is given in Equation (1.2). All regressions include control variables as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.17.: Robustness to Changes in Weather Controls (Task Complexity)

	Share Easy Issue Events (1)	*Lines added per PR* (2)	*Files changed per PR* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A.						
	0.0001^{*}	-0.0017^{**}	-0.0010^{***}	0.0021^{*}	-0.0412^{**}	-0.0224^{***}
	(.0001)	(.0007)	(.0003)	(.0012)	(.0170)	(.0083)
First Stage F-Stat.	120	88	88	90	63	63
Weather Controls	none					
Panel B.						
	0.0001^{*}	-0.0014^{*}	-0.0010^{**}	0.0027^{*}	-0.0374^{**}	-0.0244^{**}
	(.0001)	(.0008)	(.0004)	(.0014)	(.0184)	(.0096)
First Stage F-Stat.	90	66	66	68	47	47
Weather Controls	Quadratic functions of precipitation, wind speed, rel. humidity and cubic, region-specific function of mean temperature					
Panel C.						
	0.0002^{**}	-0.0012	-0.0009^{*}	0.0036^{**}	-0.0416^{*}	-0.0251^{*}
	(.0001)	(.0010)	(.0005)	(.0016)	(.0222)	(.0131)
First Stage F-Stat.	72	52	52	55	38	38
Weather Controls	Continent specific bins for precipitation, wind speed, rel. humidity, minimum and maximum temperature					
Observations	250,376	164,883	164,883	250,376	164,883	164,883

Note: The table presents IV estimates of the parameter β in Equation (1.1). Dependent variables are denoted at the top of the table. In Columns(1) to (3), the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Columns (4) to (6), an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. Relative to specifications underlying results in Table 1.2, we change the included covariates to control for weather conditions. We state the included variables at the bottom of each Panel. The first stage specification is given in Equation (1.2). All regressions include fixed effects as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.A.18.: Robustness to Changes in Fixed Effects (Task Complexity)

	Share Easy Issue Events (1)	*Lines added per PR* (2)	*Files changed per PR* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A.						
	0.0002^{**}	-0.0011	-0.0009^{*}	0.0039^{**}	-0.0368^{**}	-0.0231^{**}
	(0.0001)	(0.0009)	(0.0004)	(0.0016)	(0.0185)	(0.0112)
First Stage F-Stat.	70	50	50	54	38	38
Fixed Effects	city, region × day-of-week, region × week					
Panel B.						
	0.0002^{*}	-0.0007	-0.0006	0.0046^{***}	-0.0275	-0.0156
	(0.0001)	(0.0009)	(0.0004)	(0.0018)	(0.0187)	(0.0113)
First Stage F-Stat.	70	53	53	49	35	35
Fixed Effects	city, region × date					
Panel C.						
	0.0001^{**}	-0.0019^{**}	-0.0011^{**}	0.0033^{**}	-0.0376^{**}	-0.0183^{**}
	(0.0001)	(0.0008)	(0.0005)	(0.0013)	(0.0178)	(0.0091)
First Stage F-Stat.	85	62	62	64	45	45
Fixed Effects	city × month, region × day-of-week, region × year × month					
Panel D.						
	0.0002^{**}	-0.0017^{**}	-0.0010^{*}	0.0040^{***}	-0.0304^{*}	-0.0148
	(0.0001)	(0.0008)	(0.0006)	(0.0015)	(0.0177)	(0.0104)
First Stage F-Stat.	67	49	49	51	36	36
Fixed Effects	city × month, region × day-of-week, region × week					
Panel E.						
	0.0001^{**}	-0.0021^{**}	-0.0010^{*}	0.0029^{**}	-0.0434^{**}	-0.0196^{*}
	(0.0001)	(0.0009)	(0.0005)	(0.0012)	(0.0199)	(0.0103)
First Stage F-Stat.	96	70	70	71	49	49
Fixed Effects	region × day-of-week, city × year, city × month					
Observations	250,376	164,883	164,883	250,376	164,883	164,883

Note: The table presents IV estimates of the parameter β in Equation (1.1). In Columns(1) to (3), the regressor of interest is $PM_{2.5}$ concentration measured in $\mu g/m^3$. In Columns (4) to (6), an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. Relative to specifications underlying results in Table 1.2, we change the fixed effects. We state the included fixed effects at the bottom of each Panel. The first stage specification is given in Equation (1.2). All regressions include control variables as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. Standard errors clustered at the city level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

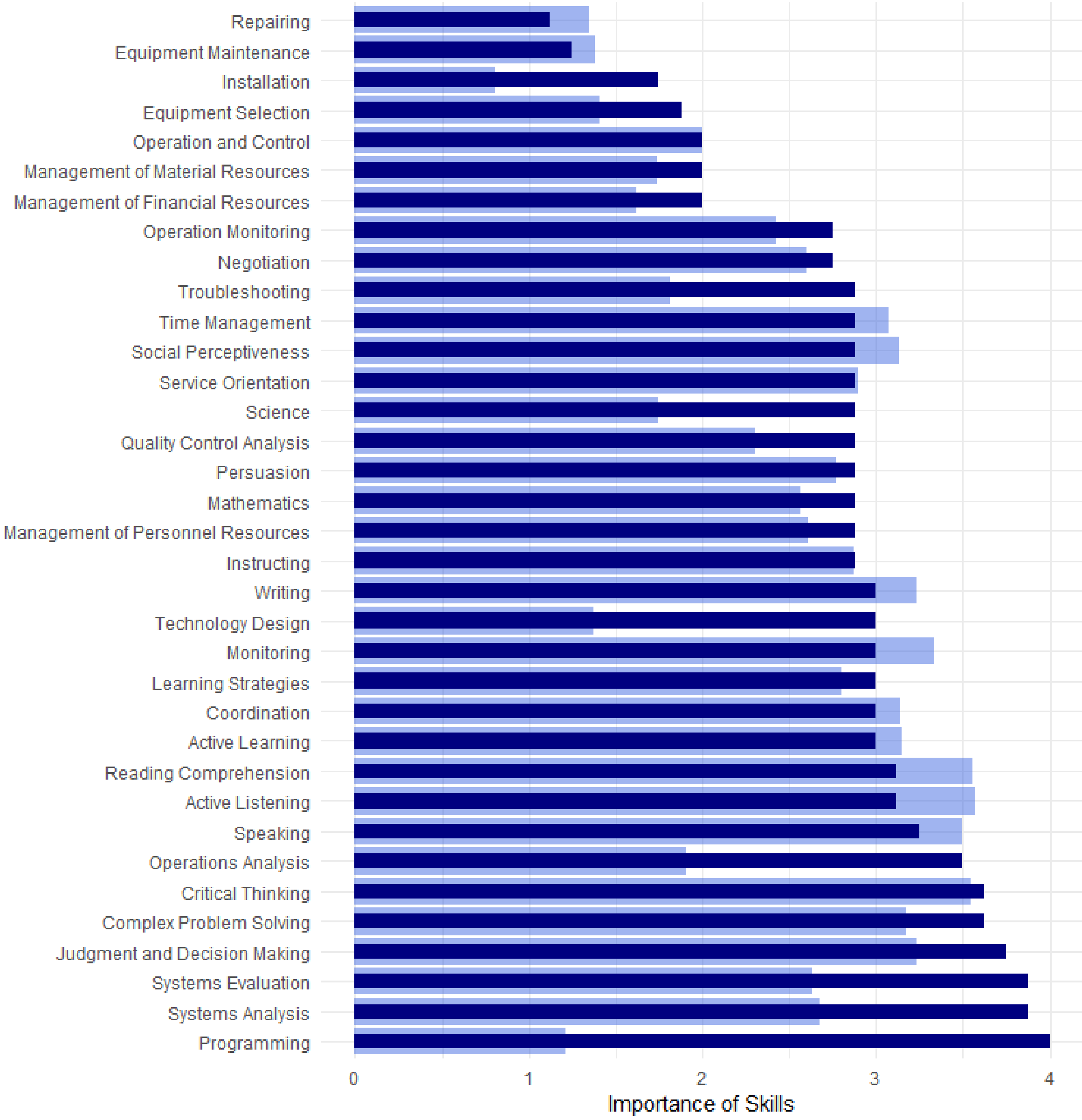
Table 1.A.19.: Effects of $PM_{2.5}$ in the Extended Sample using Inversions as IV

	Actions (1)	*Commits* (2)	*Comments* (3)	*Share Easy Issue Events* (4)	*Lines added per PR* (5)	*Files changed per PR* (6)
Panel A.						
$PM_{2.5}$	-0.0032^{***}	-0.0017^{***}	-0.0009^{***}	0.00004	-0.0023^{***}	-0.0009^{***}
	(0.0009)	(0.0006)	(0.0003)	(0.00003)	(0.0005)	(0.0003)
First Stage F-Stat.	300	300	300	301	231	231
Panel B.						
$\mathbb{1}\{PM_{2.5} > Q_{0.75}\}$	-0.1035^{**}	-0.0575^{***}	-0.0316	0.0001	-0.0383^{**}	-0.0199^{**}
	(0.0470)	(0.0196)	(0.0217)	(0.0012)	(0.0173)	(0.0092)
First Stage F-Stat.	440	440	440	385	292	292
Observations	398,687	398,687	398,687	281,985	187,935	187,935

Note: The Table presents IV estimates of the parameter β in Equation (1.1). In Panels A, the regressor of interest is $PM_{2.5}$ concentration in $\mu g/m^3$. In Panel B, an indicator for $PM_{2.5}$ concentration exceeding the city-specific 75th percentile is used instead. The excluded instruments in the first stage are interactions between a measure of inversion strength as specified in 1.3 and dummies indicating the geographic region a city is located in. Standard errors clustered at the city level are reported in parentheses. All regressions include covariates as described in Table 1.2 and are weighted by the number of active workers in a city during the current month. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

1.B. Additional Figures

Figure 1.B.1.: Skill Requirements in High-Skill Occupations and Software Development



Note: Based on data from O*NET Database Version 25.0. Skills Table. Light blue bars reflect average importance of the respective skill across all high-skill occupations, i.e. occupations in Job Zones 4 and 5. Dark blue bars reflect importance of the respective skill among software developers, i.e. occupation 15-1132.00 ("Software Developers, Applications").

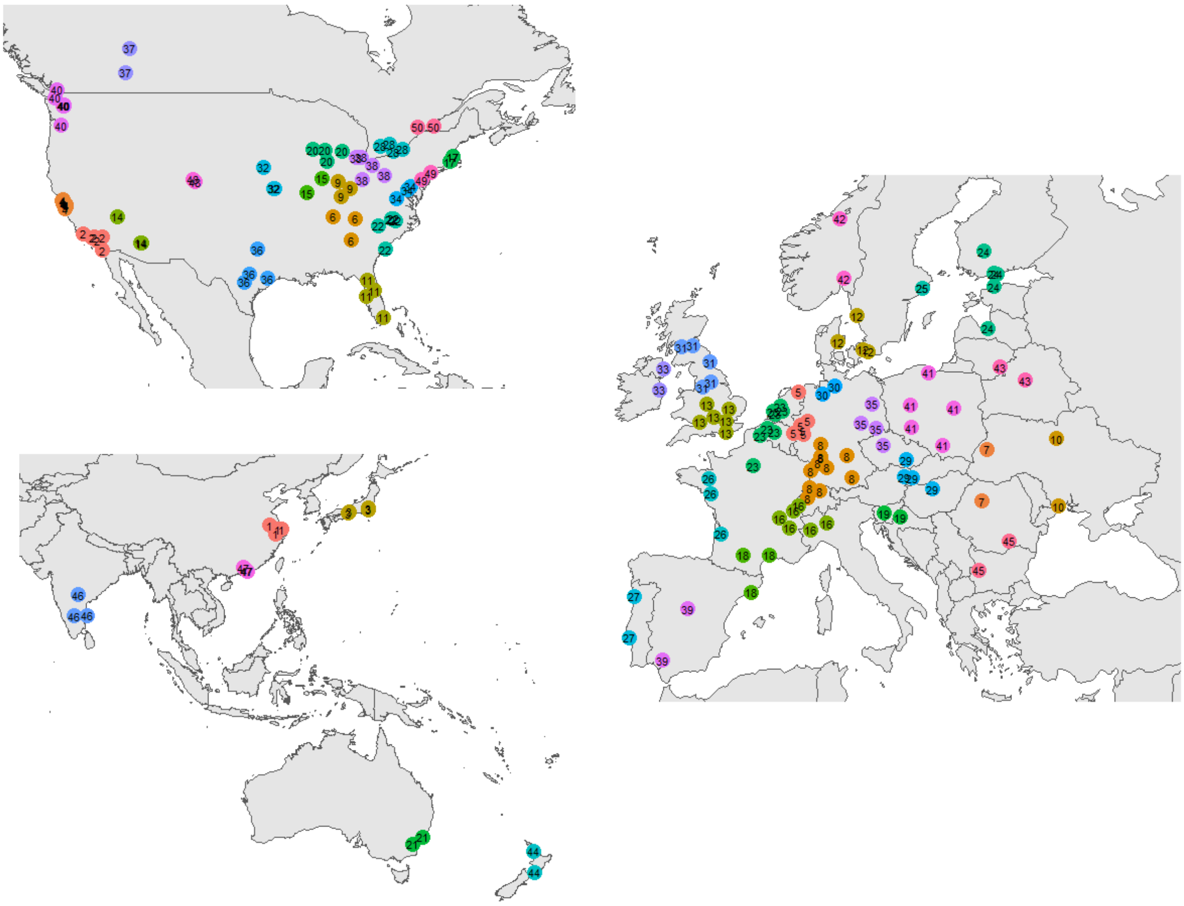


Figure 1.B.2.: Illustration of first stage city groups g

Note: Maps show our sample cities. The color of and number on top of the city markers refers to the group g we assign a city to for the first stage estimation of the effect of wind direction on air pollution (see section 1.4, especially equation (1.2)).

Figure 1.B.3.: First stage for all 50 city groups

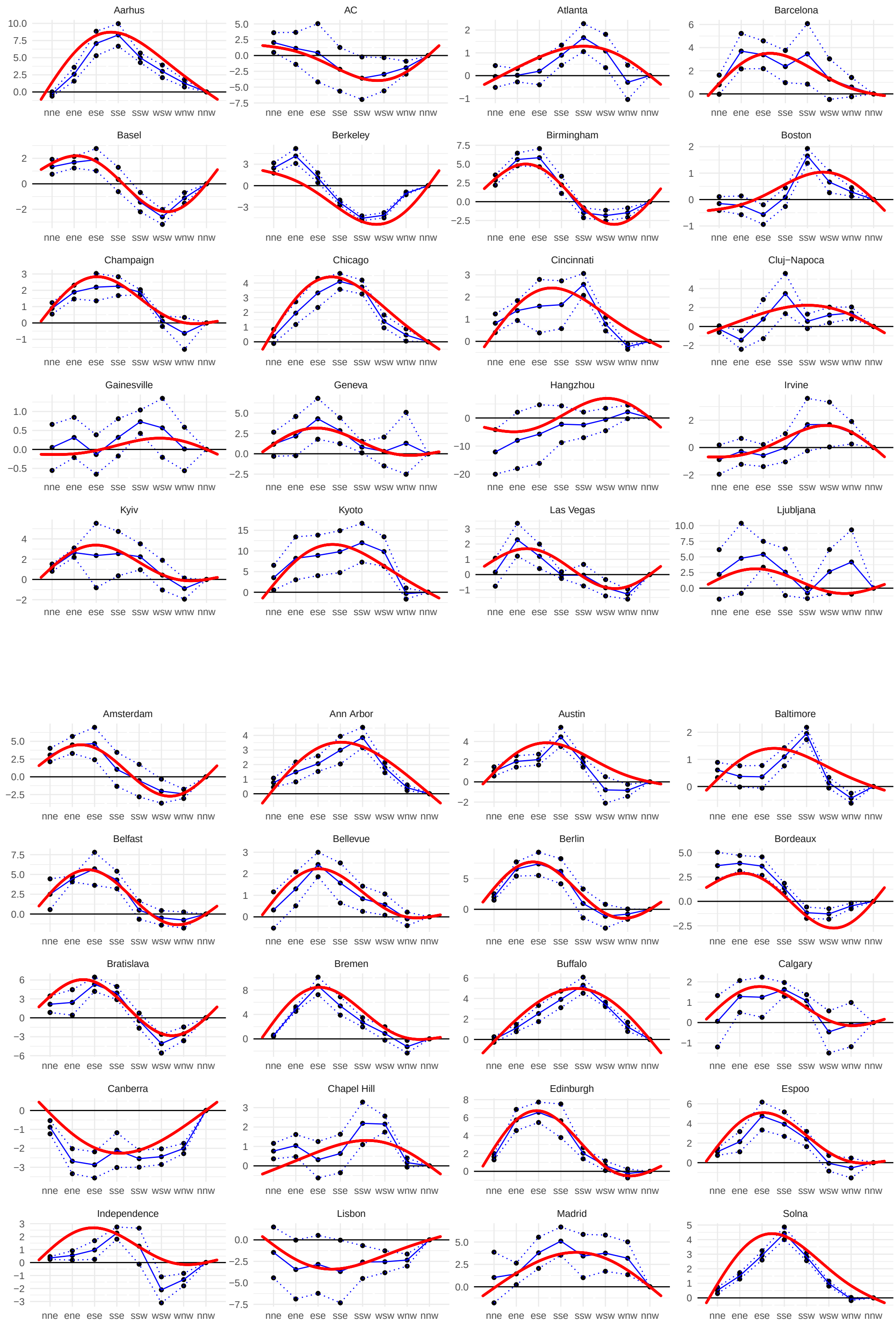
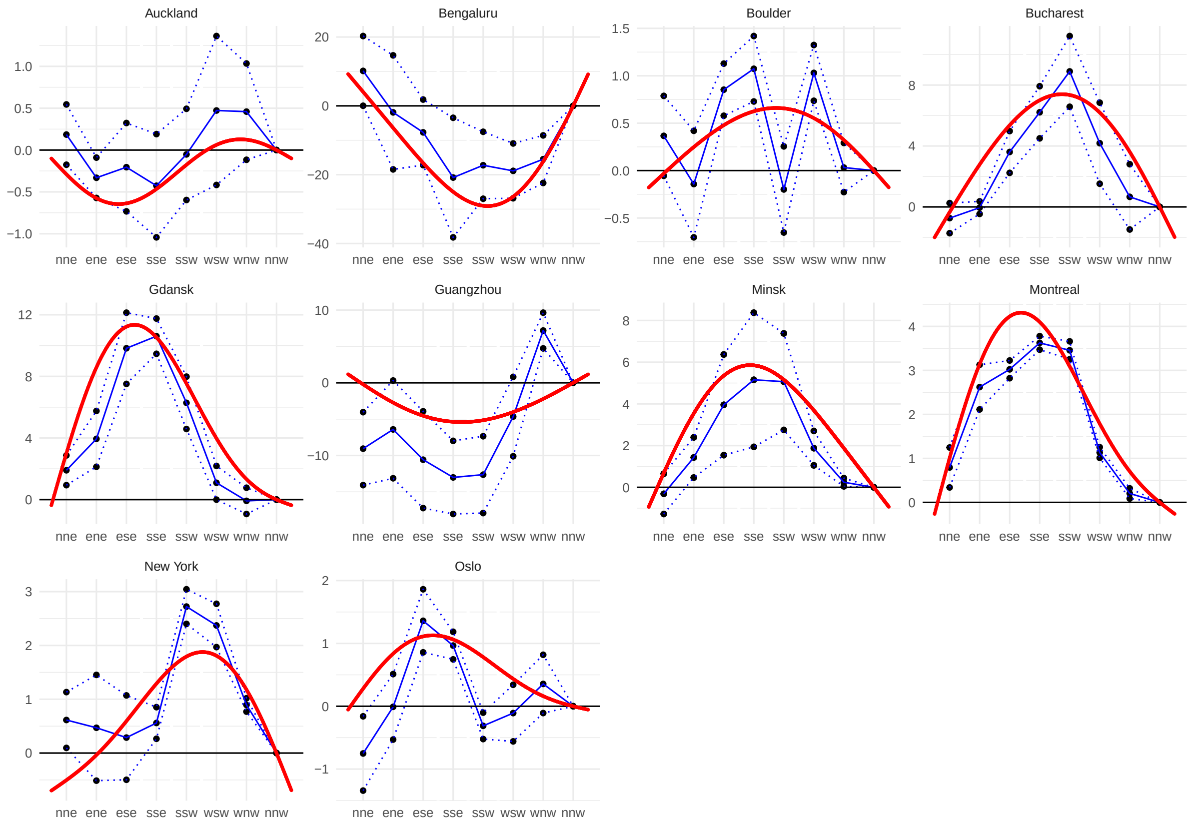


Figure 1.B.3.: First stage for all 50 city groups (continued)



Notes: Plots present estimated coefficients from regressions of $PM_{2.5}$ measured in $\mu g/m^3$ on wind direction for each first stage city group as depicted in 1.B.2. Solid blue line: connects estimated coefficients on seven dummies for seven 45° bins of wind direction. The omitted direction is north-north-west, (315°, 360°]. Dashed lines: 95% confidence intervals. Red line: estimated relationship when wind direction is parameterized as the sine of wind direction in radians and wind direction in radians divided by two. Plot titles denote one city from the respective group.

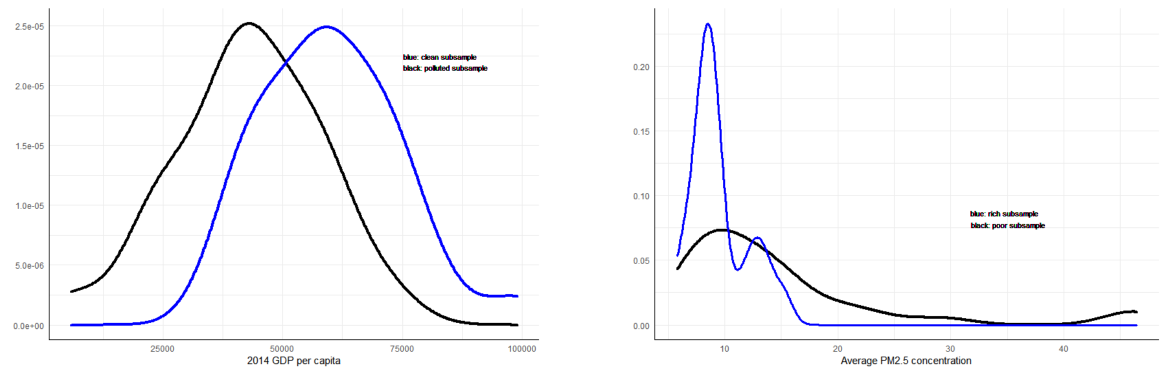


Figure 1.B.4.: GDP per capita and average $PM_{2.5}$ concentration

Note: The left plot shows the distribution of 2014 GDP per capita across city-groups g separately for the high and low pollution subsample used in the heteorgeneity analysis in Section 1.5. Blue: low pollution, black: high pollution.The right plot depicts the distribution of average $PM_{2.5}$ concentration separately for low and high income city-groups. Blue: high income, black: low income.

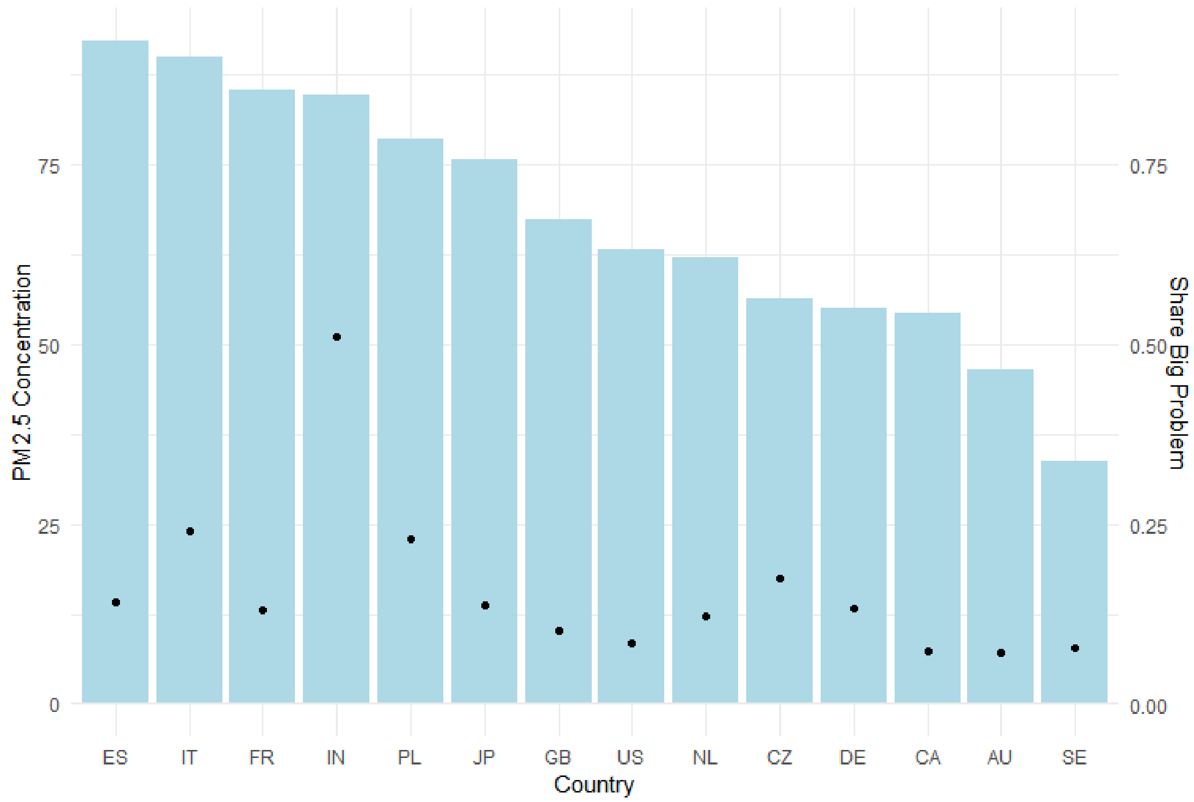


Figure 1.B.5.: $PM_{2.5}$ and Awareness by Country

Note: Bars depict the share of respondents from the respective country stating that air pollution is a big problem in their country, based on the Pew Research Center Science Survey (2020). Black dots reflect average $PM_{2.5}$ concentration in the cities within the respective country which are included in our data.

1.C. Gitcoin

This section provides additional details regarding the data collected from Gitcoin to assess the monetary value of output produced on GitHub and to validate some of our productivity and task complexity outcomes.

We collect data on 292 Gitcoin transactions via the Gitcoin API, including the type of the posted issue (bug, documentation, improvement, feature, or other), the expected issue difficulty as assessed by the issue funders (beginner, intermediate, or advanced), the URL to the PR solving the issue and awarded the payment, the value of the payment in USD, and the number of hours worked on the PR as stated by the PR author. The number of issues is relatively low compared to the volume of our GitHub data because Gitcoin is much younger than GitHub and only used by a small share of GitHub users. Using the URL of the PR, we combine this with information on pull request size obtained via the GitHub API, i.e. the number of commits it comprises, the number of lines of code added and deleted, and the number of files changed. This is possible because all Gitcoin issues and PRs are created in public GitHub repos and thus visible to us. In this context, a pull request reflects the complete work on a certain issue. Commits can be interpreted as single work steps in completing this task.

Combining the data on the amount of coding work done and on the payment made we can estimate the monetary value of output produced in public GitHub repos. The average monetary value per commit ranges from $32 in the subsample of issues of difficulty level *beginner* to $679 among issues marked as *advanced.* In the full sample, it amounts to $112. The mean time input per commit also exhibits a steep gradient with respect to difficulty: It is 1 hour at the beginner level, but 5.3 hours at the advanced level.

To validate the use of the number of commits per day as one of our core measures of developer productivity, we analyze how the number of commits in a PR correlates with the payment awarded and the time spent on it in the Gitcoin sample.

Table 1.C.1 depicts results from regressions of the payment awarded for a PR, $log(payment_i)$, on the number of commits it comprises, $commits_i$ (columns 1-3), or the logarithm thereof (columns 4-6). We run specifications without any controls (columns 1 and 4), with controls for issue difficulty, issue type and the year of PR creation (columns 2 and 5), and alternatively with repository fixed effects (columns 3 and 6). The omitted difficulty category is *advanced.* Across specifications we find statistically significant positive effects, indicating that a higher number of commits is associated with higher payments. In terms of magnitude, the results from the regressions without any controls imply that one additional commit is associated with a 5.4% increase in payment (column 1), or that a 10% increase in the number of commits is correlated with a 3.5% rise in payment (column 4). When adding controls for issue difficulty and type, the magnitude of the effect is reduced. This reduction implies that part of the increase in payments

in commits is driven by higher issue complexity. Even when using only variation across PRs submitted to the same repo, i.e., work on the same project, the positive relationship persists.

In Table 1.C.2 we present results from models where the dependent variable is $hoursworked_i$, the time input as reported by the PR author. We find that the time required to complete a task increases in the number of commits, and more so for issues of higher difficulty.

To validate our proxies for PR complexity, we run the specifications from columns 4 to 6 of Table 1.C.1 again, but add the number of files changed in the PR and the logarithm of lines of code added as additional regressors. Results are presented in Table 1.C.3. Holding the number of commits constant, adding more lines of code and changing more files is associated with a higher payment, suggesting that these variables indeed reflect task complexity.

Table 1.C.1.: Validity Check: Number of Commits and Gitcoin Payments

	Dependent variable: $log(payment_i)$					
	(1)	(2)	(3)	(4)	(5)	(6)
$commits_i$	0.054***	0.039***	0.034***			
	(.010)	(.009)	(.010)			
$\log(commits_i)$				0.348***	0.264***	0.192***
				(.071)	(.068)	(.059)
$\mathbb{1}\{Difficulty_i = Beginner\}$		−2.399***			−2.412***	
		(.439)			(.419)	
$\mathbb{1}\{Difficulty_i = Intermediate\}$		−1.878***			−1.851***	
		(.415)			(.405)	
Year dummies		✓	✓		✓	✓
Issue difficulty		✓			✓	
Issue type		✓			✓	
Repository fixed effects			✓			✓
Observations	292	274	292	292	274	292

Note The table presents results from OLS regressions using data on the sample of Gitcoin pull requests. Observations are at the pull request level. Dependent variable is the logarithm of the payment awarded to the PR author. Explanatory variables are the number of commits (columns 1 to 3) or the logarithm thereof (columns 4 to 6). Columns 2 and 5 add dummies for the year the pull request was created, dummies for issue difficulty, and dummies for issue type. Column 3 and 6 instead add dummies for the year the pull request was created and fixed effects for the repository. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.C.2.: Validity Check: Number of Commits and Hours Worked on a PR

	hours worked$_i$			
	(1)	(2)	(3)	(4)
$commits_i$	0.375***	0.939***		
	(.132)	(.341)		
$\log(commits_i)$			2.375***	10.346***
			(.648)	(3.535)
$\times\ \mathbb{1}\{Difficulty_i = Beginner\}$		−0.882**		−9.748***
		(.354)		(3.574)
$\times\ \mathbb{1}\{Difficulty_i = Intermediate\}$		−0.667*		−8.471**
		(.349)		(3.560)
Observations	271	267	271	267

Note The table presents results from OLS regressions using data on the sample of Gitcoin pull requests. Observations are at the pull request level. Dependent variable is the number of hours worked reported by the PR author. In colunm 1 the only explanatory variable is the number of commits in the PR. Column 2 adds dummies for issue difficulty and interactions between the number of commits and the difficulty dummies. The ommited difficulty category is *advanced*. In columns 3 and 4 report results from the same models except that the number of commits is replaced by the logarithm thereof. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 1.C.3.: Validity check: PR complexity and Gitcoin payments

	$log(payment_i)$			
	(1)	(2)	(3)	(4)
$log(commits_i)$	0.143**	0.136**	0.070	0.145**
	(0.067)	(0.068)	(0.058)	(0.056)
$files\ changed_i$	0.005	0.007*	0.011***	0.004
	(0.005)	(0.004)	(0.004)	(0.004)
$log(lines\ added_i)$	0.152***	0.112***	0.091***	0.150***
	(0.036)	(0.035)	(0.028)	(0.033)
$easy\ label_i$				−0.348**
				(0.173)
Year dummies	✓	✓	✓	✓
Issue difficulty dummies		✓		
Issue type dummies		✓		
Repository fixed effects			✓	
Observations	292	274	292	270

Note The table presents results from OLS regressions using data on the sample of Gitcoin pull requests. Observations are at the pull request level. Dependent variable is the logarithm of the payment awarded to the PR author. Explanatory variables are the number of commits and the number of lines of code added in the PR (both in logs), the number of code files changed and dummies for the year the pull request was created. Column 2 adds dummies for issue difficulty and issue type. Column 3 instead adds fixed effects for the repository. Column 4 instead adds a dummy variable taking a value of one if the issue addressed by the PR carries a label that we classify as indicating an easy issue. The number of lines of code added and of files changed in the PR are winsorized at the 1st and the 99th percentile. Robust standard errors are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

1.D. Auxiliary Regressions

For estimating equation (1.1), measures of the output of each individual developer i are aggregated to the city-day level. Instead of forming simple averages, we take into account additional information at the developer level. This is done by estimating auxiliary regression, a common approach in this literature (e.g. Currie et al., 2015). In a first step, we estimate regressions for outcome y of developer i living in city c on day d of the following kind.

$$y_{i,c,d} = \mu_i + \mathbf{x}'_{i,d}\pi + \psi_{c,d} + \varepsilon_{i,c,d} \tag{1.1}$$

Here, $y_{i,c,d}$ denotes one of the measures of developer output, task choice, or working hours. The fixed effect μ_i captures time-invariant unobserved factors at the developer level. Including these is important as the composition of developers changes over time. A developer's experience is controlled for by $\mathbf{x}_{i,t}$, a vector of indicators for time since registration on GitHub, where each indicator represents a time span of three months. Additionally, equation (1.1) includes city-day fixed effects. Their estimates $\widehat{\psi}_{c,d}$ give the average outcome for a city-day after controlling for experience and composition effects. These estimates replace the dependent variable in equation (1.1).

This approach is computationally less costly and asymptotically equivalent to directly estimating the regressions at the individual developer level (Donald and Lang, 2007). We take into account the sample variance by weighting all regressions by the number of underlying developer observations in each city–day cell (cf. Currie and Neidell, 2005; Isen et al., 2017a).

2. Prenatal Exposure to Air Pollution and the Development of Noncognitive Skills

2.1. Introduction

Air pollution imposes high costs on society since it adversely affects several dimensions of human health and well-being. Poor air quality is also an obstacle to social mobility: Isen et al. (2017b) find that exposure to air pollution while in utero and during the first year of life reduces earnings and employment during adulthood. Voorheis (2017) confirms this result and finds that educational outcomes are negatively affected as well, while recent evidence by Colmer and Voorheis (2020) suggests that these adverse impacts even extend to the next generation. Since low-income families and minorities often live in more polluted neighbourhoods than more affluent groups (see e.g. Banzhaf et al., 2019; Currie et al., 2023; Glatter-Götz et al., 2019; Rüttenauer, 2018, for evidence on the US and Europe), these long-run effects of gestational pollution exposure not only impose a substantial economic cost on society, but also inhibit equality of opportunity.

Optimal policy responses to this issue might depend on the mechanisms driving the adverse long-run effects of early pollution exposure. Educational achievement and and labor market success are functions of human capital, whose core components are cognitive and non-cognitive skills.[1] While the predictive power of the two types of skills for educational attainment and labour market performance is comparable, they differ crucially in how they respond to intervention programs and investments: There is growing evidence that non-cognitive skills are malleable up until adulthood and can be improved by way of low-cost interventions implemented in the classroom- or even work-environment (Adhvaryu et al., 2018; Alan et al., 2019; Sorrenti et al., 2020), whereas cognition is less malleable, especially after school start

[1]While cognitive ability captures intelligence, the ability to reason and understand complex ideas, non-cognitive skills - also known under names such as socio-emotional skills, soft skills, or personality traits - comprise a variety of abilities that are weakly correlated with intelligence, such as social competencies, emotional stability and persistence.

age (e.g. Almlund et al., 2011; Cunha et al., 2010). Hence, understanding which mechanisms drive adverse long-term effects of early-life pollution exposure, and how important the respective channels are in quantitative terms, is paramount when deciding about feasible and appropriate policy responses: If long-term effects were driven purely by reduced cognitive skills, the only option to avoid them would be to reduce air pollution. If, on the other hand, non-cognitive abilities play a relevant role as well, long-term effects can also be alleviated ex-post via intervention programs or investments targeting these abilities. Given that it can be extremely costly or even impossible to reduce pollution levels in some circumstances (e.g. pollution arising from another jurisdiction via transboundary atmospheric transport or from natural sources like wildfires), investigating whether alternative options exist to compensate individuals ex-post for exposure to high levels of pollution while in utero, is highly policy-relevant.

The existing evidence regarding these channels is incomplete. A number of studies find that prenatal pollution exposure causes worse performance in standardized achievement tests taken in primary and high school (Bharadwaj et al., 2017; Sanders, 2012), as well as tests of fluid intelligence Molina (2021), pointing strongly towards cognitive ability as a relevant channel.[2] Regarding the second main component of human capital, non-cognitive abilities, causal evidence is missing. Therefore, the aim of this paper is to answer the question whether in-utero exposure to air pollution has a causal impact on non-cognitive abilities, and to assess how important this potential channel is, relative to the cognitive ability mechanism.

I employ data on non-cognitive abilities during childhood from the German Socio-Economic Panel (Goebel et al., 2019). Specifically, the survey includes mother-reported Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism) assessed at ages 2-10 for children born between 2000 and 2015. I combine this with data on particulate matter with a diameter of less than 10 μm (PM10) measured at outdoor monitors and reanalysis data on meteorological conditions. To address the issues of endogeneity and measurement error in particulate matter exposure, I exploit plausibly exogenous variation in thermal inversions, a meteorological phenomenon that deteriorates air quality (following e.g. Arceo et al., 2016; Jans et al., 2018b; Molina, 2021).

Results show that prenatal pollution exposure raises neuroticism, or put differently, reduces emotional stability. A 1 unit increase in gestational PM10 exposure raises Neuroticism measured at ages 5 to 10 by approximately 7% of a standard deviation. The effect is mainly driven by exposure during the second and third trimester of pregnancy, and by increases in Neuroti-

[2] In fact, while IQ and school achievement test scores are commonly considered as measures of cognitive ability, certain personality traits, esp. Conscientiousness and Emotional Stability, were found to have predictive power for these outcomes as well. However, cognitive skills can typically explain a larger part in the variation of these outcomes. (e.g. Almlund et al., 2011; Borghans et al., 2008; Moutafi et al., 2006)

cism at the upper part of the distribution. Other dimensions of the Big Five are not affected. The effect on Neuroticism is of the same order of magnitude as the impact on measures of cognitive ability found in earlier work. Since existing research established a negative correlation between Neuroticism and labour market outcomes, it is a plausible channel underlying the long-run adverse effects of early-life pollution exposure. Back-of-the envelope computations imply that an increase of PM10 by one standard deviation could cause reductions in earnings by roughly $0.24 - 0.29\%$ via the increase in Neuroticism.

This study contributes to the literature on the long-run consequences of early-life exposure to air pollution by providing causal evidence for a new relevant outcome, namely non-cognitive abilities. Using data on cohorts born in Germany after 2000, I study a setting with relatively low baseline pollution levels, similar to concentrations prevailing today in many developed countries.[3] I thus shed light on a different part of the dose-response function than existing papers on the effects of gestational pollution exposure on skills, educational or labour market outcomes which are based on data from developing countries (Bharadwaj et al., 2017; Molina, 2021; Rosales-Rueda and Triyana, 2019) or cohorts born in the US during the 1970s-1980s (Isen et al., 2017b; Sanders, 2012; Voorheis, 2017), i.e. settings with much higher air pollution levels. Most closely related in terms of the analyzed outcome is a study that was conducted independently of and in parallel to this dissertation by Ai et al. (2023) who find adverse impacts of in-utero exposure to PM2.5 on mental health at age 10-15 using survey data from China and average wind speed during the year of birth as instrumental variable. It is unclear, however, whether wind speed satisfies the exclusion restriction because it might be notable and trigger behavioral responses among pregnant mothers. Moreover, while mental health is closely related to some components of non-cognitive ability, I analyze a broader range of non-cognitive skills as outcomes and consider a very different context of a developed country with relatively low pollution levels.

The paper also adds to the literature on the development of non-cognitive skills. Related studies e.g. analyze the effects of family income (Akee et al., 2018), birth order (Black et al., 2018), parents' labor market incentives (Hufe, 2020) or child care arrangements (Datta Gupta and Simonsen, 2010). Persson and Rossin-Slater (2018) and Adhvaryu et al. (2019) also investigate the impact of prenatal conditions, specifically maternal stress and malnutrition, on later life mental health and non-cognitive ability. Most closely related is work by Grönqvist et al. (2020) who analyze how childhood exposure to lead affects adult human capital and crime, and identify negative effects non-cognitive skills as an important mechanism. Relative to this study, I focus on less toxic air pollutants, which - as an externality of economic production and traffic - are omnipresent in both developing and developed countries. Thus, I complement the

[3] Mean gestational PM10 exposure in my sample is approximately 24 µg/m^3.

existing literature by analyzing the role of this pertinent environmental factor in the formation of non-cognitive skills.

The remainder of the paper is structured as follows: The next section provides background information on prenatal pollution exposure and non-cognitive abilities. Section 3 describes the data used in the analysis. I present the empirical specification in section 4 and the results in section 5. Section 6 concludes.

2.2. Background: Noncognitive Skills and Prenatal Air Pollution Exposure

Noncognitive skills are important predictors of educational achievement and labor market outcomes, even beyond their impact on education (Heckman et al., 2006; Lindqvist and Vestman, 2011), and over recent decades the returns to these skills increased relative to the returns to cognitive ability (Deming, 2017; Edin et al., 2017). To capture the different dimensions of noncognitive ability, I rely on the Big Five personality traits, assessed during childhood, as outcome variables. The Big Five are a widely used taxonomy which was developed in personality psychology. Among the five traits, conscientiousness (the tendency to be organized, responsible, and hard-working) and emotional stability (the opposite of neuroticism) show the most robust positive correlations with labor market success (e.g. Almlund et al., 2011; Cubel et al., 2016; Fletcher, 2013). Thus, these two traits in particular are potential mediators for the documented long-run effects of pollution exposure. Focusing on noncognitive ability during childhood is based on ample evidence in developmental psychology showing that by school start age children already differ substantially in character traits and that these are predictive of their adult personality (e.g. Almlund et al., 2011; De Pauw, 2017; Deal et al., 2005). Moreover, childhood noncognitive ability is a significant determinant of school performance and educational outcomes (e.g. Carneiro et al., 2007; Currie and Stabile, 2006; Johnston et al., 2014).

Findings from brain lesion studies, neuroimaging and psychopharmacological research imply that the source of all noncognitive abilities lies in the brain (e.g. Almlund et al., 2011). Personality neuroscience uses neuroimaging techniques to identify how personality traits depend on brain structure and function as well as levels of hormones and neurotransmitters. While the field is relatively new and the high costs of neuroimaging often restrict sample sizes, some findings already emerged as relatively robust, e.g. the association of Extraversion with dopamine or the association of Neuroticism with cortisol and activity in the hippocampus (Allen and DeYoung, 2017). As skills that depend on the functioning of the brain, noncognitive abilities might plausibly be affected by exposure to air pollution, as the latter not only causes respiratory and cardiovascular diseases, but can also induce damage to the central nervous system. The medical literature shows that very small particles reach the brain tissue where they cause oxidative stress and neuroinflammation. Since brain formation and growth proceed very rapidly during the prenatal period, air pollution exposure during this critical time window might cause irreversible damages to the nervous system by disrupting these processes (de Prado Bert et al., 2018). The pollutants that are most likely to cause permanent reductions in cognitive and noncognitive ability in this way are carbon monoxide (CO) and ultrafine particles, as both can cross the placenta, and thus pose most harm to the foetus. Possible mechanisms are maternal systemic and placental oxidative stress and inflammation and impaired transport of oxygen and nutrients to the fetus (Levy, 2015; Johnson et al., 2021). Recent evidence from brain imaging indeed suggests that prenatal air pollution exposure is associated with a reduction in white matter volume and changes to brain structure in humans (de Prado Bert et al., 2018; Beckwith et al., 2020), however given very small sample sizes and non-random variation in pollution exposure, such associations cannot be interpreted as reflecting causal effects.[4]

Given these premises, several epidemiological studies investigate the correlation between in-utero or early childhood exposure to air pollution and mental health issues (e.g. ADHD or Autism) or behavioral problems in childhood, i.e. outcomes related to noncognitive ability.[5] Annavarapu and Kathi (2016), Myhre et al. (2018) and Xu et al. (2016) provide reviews of this literature and point out common caveats: Most studies are based on cross-sectional comparisons between individuals living in different places, and thus prone to omitted variable bias. Hence, while suggestive of a relationship between early air pollution exposure and noncognitive ability, the results of this literature do not reflect causal effects.

2.3. Data

2.3.1. Socio-economic Panel: Noncognitive skills

Estimating the impact of early life pollution exposure on the formation of noncognitive ability requires data that not only includes information on individuals' socioemotional skills but also on their location and time of birth. Besides, the data must cover a sufficiently large number of individuals from multiple birth cohorts, since only temporal variation in particulate matter and inversions is used in the research design. Many surveys that contain extensive information about noncognitive ability do not meet these requirements as they cover only single birth co-

[4]For complementary evidence from animal studies using random variation in pollution exposure see e.g. Woodward et al. (2018) or Costa et al. (2014) for a review.

[5]Mental health issues are often interpreted as extreme realizations of personality traits (Almlund et al., 2011; Widiger et al., 2017) and, especially at child age, frequently measured by the same concepts as noncognitive ability (Currie and Almond, 2011; Johnston et al., 2014).

horts, e.g. the British longitudinal cohort studies or the studies from the U.S. Early Childhood Longitudinal Study program. Other data sets lack the necessary disaggregated geographical information, such that individual pollution exposure cannot be determined. Lastly, frequently used administrative data sources on cognitive and noncognitive skills, e.g. the Swedish military enlistment data, cover individuals who were born in a place and at a time for which accurate measurements of air pollution are unavailable.[6]

A data source that addresses most of these demands is the German Socio-Economic Panel (SOEP), a large household panel survey started in 1984, which covers roughly 15,000 households and 30,000 individuals (SOEP, 2019). It includes mother-reported Big Five personality traits for all children aged 2-10 in SOEP households which I use as measures of childhood noncognitive abilities. The relevant information is based on the "Mother-and-child" questionnaires which were introduced in 2005 for 2- to 3-year-old children, in 2008 for 5- to 6-year-olds and in 2012 for 9- to 10-year-olds. Mothers answer questions on their child's behavior on a scale from 0 to 10. Each question can be mapped into one of the Big Five domains. Mothers' of 2-3 year old children are asked to answer only one question per trait and neuroticism is not yet included. When children are in the older two age groups, two questions are included per domain and all five traits are assessed. The questions are presented in Appendix Table 2.A.1. One important thing to note about the questionnaires is that the items intended to measure openness are likely to capture at least in part the child's cognitive ability (e.g. '*My child is quick at learning new things vs. needs more time*'). Thus, the main focus of my analysis will lie on the other four traits which isolate non-cognitive skills. To construct the outcomes of interest I collect the relevant information from the 2005-2018 mother-and-child questionnaires, I recode items where necessary to ensure that high values reflect higher realizations of the respective trait, add up values for items within each domain and standardize the resulting scores within each age group (2-3, 5-6 and 9-10). For children observed at multiple ages, I keep only the most recent observation, as personality differences become more pronounced with age and my main interest is in long-run effects of pollution exposure. The relevant information is available for individuals born between 2000 and 2012 for Neuroticism (*Sample II*) and for those born 2000-2015 for the other four traits (*Sample I*), i.e. the included birth cohorts extend over more than a decade.

To assign pollution exposure to individuals, I rely on information on year and month of birth as well as county of residence. For the majority of children (71% in *Sample I* and 65% in *Sample II*), I can identify the county of birth, as the county of residence during the year

[6]The enlistment data offers a large sample size and high quality measures of noncognitive skills based on interviews with a trained psychologist, but is not suitable for my analysis, because air quality data for Sweden is not available at a large scale before 1980. Due to falling demand for conscripts, the share of a birth cohort that was enlisted fell to roughly 70% for mid-1980 born cohorts, implying potential selection bias problems when analyzing cohorts born after 1980. (Grönqvist et al., 2017)

of birth. In the remaining cases, the households entered the panel after the child was already born such that the place of birth is unknown. As a proxy for county of birth, I assign the county of residence during the first wave the household was interviewed. Wrong assignments can induce measurement error in early-life particulate matter exposure which is not addressed by the instrumental variable strategy. Since measurement error causes attenuation bias, any results would reflect a lower bound. I restrict the sample to children born in a county with at least 20 observations in my sample since effects are identified from variation in environmental conditions during early life between children born within the same county. This restriction is intended to ensure that the comparison groups are large enough to avoid spurious results generated by outliers. This yields 9,470 individuals across 192 counties in *Sample I* and 6,548 individuals across 155 counties in *Sample II*. Lastly, the SOEP provides a multitude of relevant demographic and socioeconomic background variables, e.g. the child's gender, age in month and migration history, whether it lives in a single-parent household, parental education, and the number of siblings.[7] Basic sample characteristics are summarized in the middle part of Table 2.1.

Table 2.1.: Summary Statistics

	Sample I (B5 \ Neuro)	*Sample II* (Neuroticism)
Observations	9,470	6,548
Counties	192	155
Years of Birth	2000-2015	2000-2012
Age	6.9 (2.9)	8.2 (2.0)
Migration Background [%]	31.7	29.7
College-educated mothers [%]	26.8	26.7
Single-parent households [%]	15.6	17.6
PM10 in-utero [$\frac{\mu g}{m^3}$]	24.1 (5.0)	24.6 (5.1)
NO2 in-utero [$\frac{\mu g}{m^3}$]	28.8 (10.6)	29.0 (10.8)
CO in-utero [$\frac{\mu g}{m^3}$]	460.2 (193.3)	480.7 (203.5)
Inversions in-utero [%]	39.9 (7.1)	39.7 (7.0)

Note: Summary statistics based on data from the SOEP, version 35. The table reports mean values and standard deviations (in parentheses).

Given that the Five Factor Model is the most common taxonomy of personality and widely

[7] To measure parental education I construct dummy variables reflecting a low, medium or high level of education for mothers and fathers. Low education is defined as less than high school (Abitur) or vocational training, medium education is defined as having completed high school or vocational training, but no tertiary degree, and high education is defined as having completed a tertiary degree. I also include children in the sample when information on parental education is missing, and define separate dummies for these cases.

employed in economics, using the Big Five personality traits as main outcomes allows to benchmark my results against other studies and to conduct back-of-the-envelope calculations translating effects on noncognitive abilities into effects on earnings. However, the fact that these variables are based on maternal assessments and short scales with only one or two items per domain might raise concerns about measurement error. Thus, in the following I present some descriptive evidence to illustrate that these measures do contain substantial information regarding the children's non-cognitive skills.

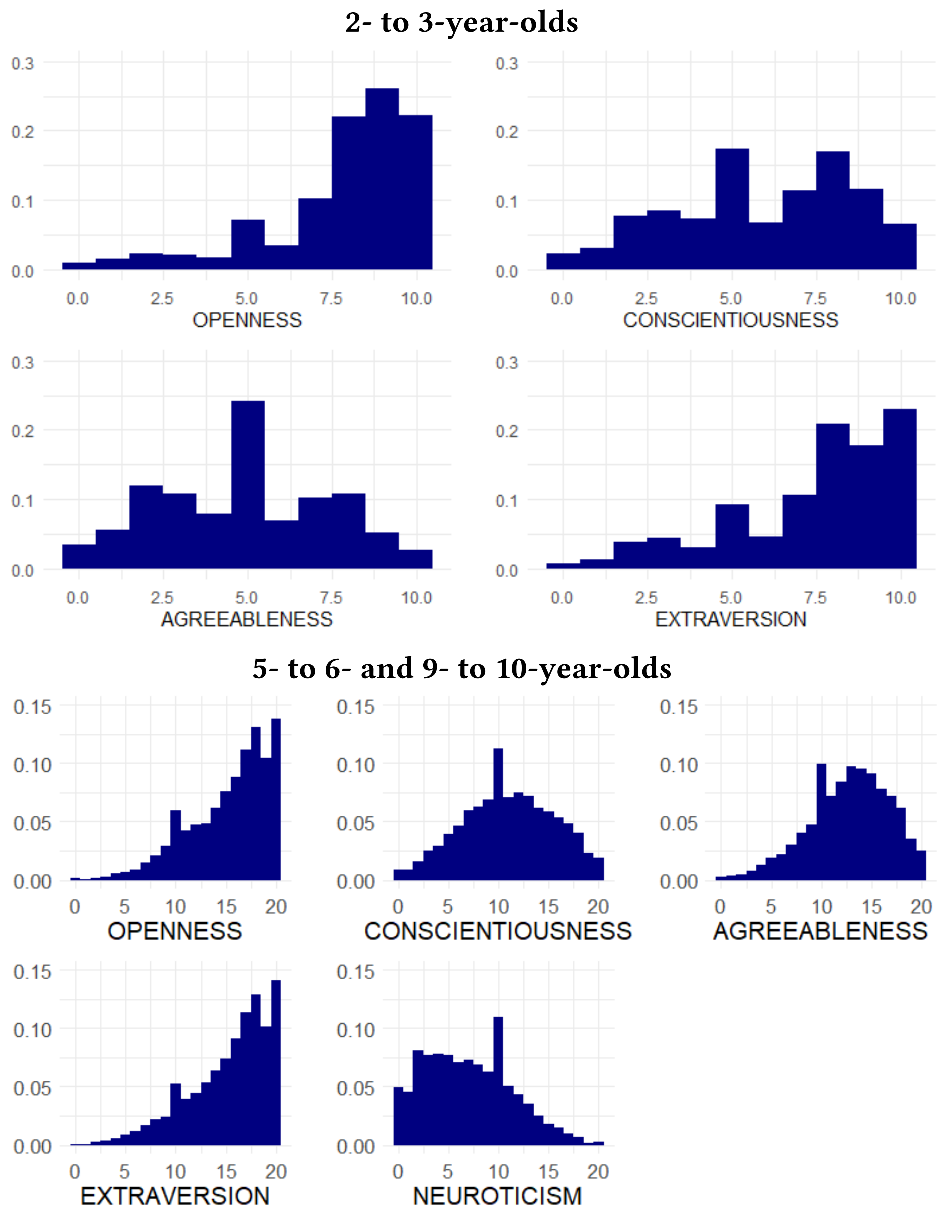


Figure 2.1.: Distribution of Big Five personality traits by age group.

Note: The distributions are based on all available observations in the SOEP, i.e. including multiple observations for the same child and children born in counties with less than 20 individuals. Distributions among 2- to 3-year-olds are based on 8,037 observations, distributions among older children on 11,657 observations.

Firstly, Figure 2.1 shows the distribution of the Big Five traits by age group (not standardized). The upper four plots show distributions for the age range from 2-3 years. While there is some clustering at high values for Openness and Extraversion which are picked by 20% to 25% of all mothers, all potential values do occur with positive frequency across all four outcomes. In the older age groups, the distributions exhibit even more variation as mothers answer two instead of one question per domain (bottom five plots). One important thing to note is that the distributions for all five traits exhibit a spike at the intermediate value of 10, with different intensity across outcomes. A potential explanation for this is that some mothers have difficulties when trying to assess their children's noncognitive skills and thus opt for the middle values. However, the vast majority of respondents do not follow that strategy, suggesting that their answers are informative about the personality of their children. In Appendix Figure 2.B.1, the distributions are depicted by maternal education. The variables exhibit strong variation in all socio-economic groups, but less educated mothers tend to choose intermediate values slightly more often than highly educated mothers. Given that I control for parental education in the regressions and conduct falsification tests showing that the instrument is not significantly correlated with family characteristics, this should not affect the results. In the main analysis, I thus use the data as it is, but in a robustness check, I drop observations where mothers opted for the intermediate values across several items.

To further explore the validy of the outcome variables, I show correlations of the mother-reported Big Five with father-, and self-reported personality traits in Figure 2.2.[8] Correlations are measured between the mother-reported Big Five assessed at ages 9-10 and (i) father-reported Big Five assessed at the same age range, (ii) self-reported Big Five at ages 11-12 and (iii) self-reported Big Five at ages 16-17. All variables are standardized to have mean zero and standard deviation one within the respective age-by-respondent cell. All correlations are positive and statistically significant, implying not only that mother-reports are in line with reports by others for the same age range, but also predictive for noncognitive ability during adolescence, at an age relatively close to transition into the labor market or higher education.

In Appendix Figure 2.B.2, I also show partial correlations between the mother-reported Big Five and measures of the child's school performance, well-being and preferences. In each case I control for parental education, a single-parent household dummy, child gender and migration background. Firstly, mother-reported Big Five when the child is aged 5-6 years are relevant predictors of the mother-assessed probability that the child will graduate from the academic track of the German school system when the child is 7 to 8 years old (i.e. before

[8] Father-reported Big Five are available only for a subsample of 1,144 children at age 9-10. Self-reported Big Five are available from age 11-12 onward. Most of the adolescents observed at these ages live in households who were not yet part of the SOEP during their year of birth. For these reasons, I rely solely on the mother-reported Big Five in the main analysis.

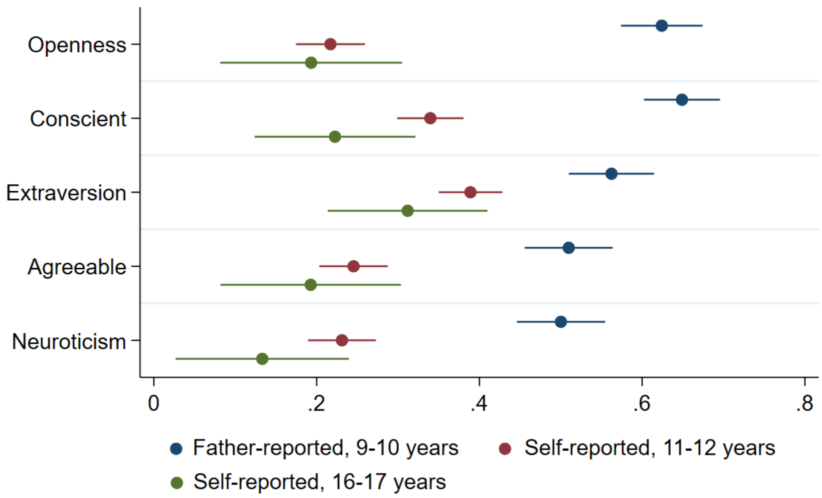


Figure 2.2.: Mother-, father- and self-reported Big Five

Note: The figure presents correlations between mother-assessed and either father- or self-reported Big Five personality traits. Father-reported Big Five are measured with the same scales as mother-reported variables based on sample sizes ranging from 1,085 to 1,097 depending on the domain. Self-reported Big Five at ages 11-12 and 16-17 are measured with three items per domain and available for 2,215 to 2,266 and 355 to 361 individuals, respectively who were also observed at age 9-10. 95%-confidence intervalls are based on heteroscedasticity robust standard errors.

track choice is made). Openness is a strong positive predictor, which is unsurprising, given that openness as measured in the SOEP likely reflects cognitive skills to some extent. More importantly, conscientiousness and neuroticism are also relevant predictors, with positive and negative sign, respectively. Child-reported life satisfaction and risk aversion at ages 11-12 are both correlated with mother-assessed Big Five at age 9-10: Agreeableness, Extraversion and Conscientiousness are positively associated with life satisfaction. All five traits are correlated with risk aversion, with the strongest, positive association for Neuroticism. The signs of these correlations are plausible, providing further support for the validity of the mother-reported noncognitive skills.

2.3.2. Pollution, Weather and Inversions

I obtain data on daily PM10 concentration measured at outdoor monitors between 2001 and 2016 from the federal environmental agency (Umweltbundesamt [UBA]). I use data from 181 stations which were active throughout the full time period. A map of the monitor locations is provided in Appendix Figure 2.B.3. To assign PM10 to counties, I use inverse distance weighting, based on stations within a radius of 60km around the county centroid.[9] The median distance between the county centroid and the assigned stations is 25.8 km. I aggregate daily

[9] I restrict the maximum number of stations per county to three since using more stations increases the number of missing values. In a small number of large cities, which are also counties, there are more than three monitors within the city boundaries. In these cases I assign all monitors within at most 20km distance.

PM10 concentrations to trimesters (= 90 days) and in-utero periods (= 270 days). I keep only observations with less than 115 missing daily values during the in-utero period and at most 40 missing values per trimester. Given that the European Union only introduced binding limit values for PM10 in 2005, there were relatively few measurement stations in Germany during the early 2000s, with the number of monitors rising sharply only in 2003. Hence, the dataset does not cover the full sample period (the oldest individual in the sample were born in 2000) and only 190 out of the 192 relevant counties. On top of that, there is a non-trivial number of missing measurements in the data. In total, for 14.5% of the final sample, the in-utero period PM10 concentration is missing. Hence, I will report both 2SLS and reduced form results, since the first are informative about the quantitative effect of pollution on noncognitive abilities, but the second are based on a larger sample. I also collect data on nitrogen dioxide (NO2) and carbon monoxide (CO) concentration from UBA. They are common co-pollutants of PM10, and CO can also cause damage to the unborn child.[10] I proceed in the same way as described above to transform monitor-by-day observations into county level concentrations during the periods of interest.

To construct the instrumental variable based on thermal inversion periods, I employ reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) on surface level and upper air temperature for the years 1999-2016. During normal times, air temperature decreases with altitude. A thermal inversion occurs when the relationship between temperature at different altitude levels is reversed, i.e. temperature increases with altitude. The data are available at an hourly level on a regular 0.25° latitude x 0.25° longitude grid.[11] Upper air temperature is measured as temperature at a pressure level 50 hPa below the surface level pressure in the county, which corresponds to approximately 400-500m higher altitude. I then assign data from grid points to counties, based on the inverse distance weighting method, including all grid points within a 30km radius around a county centroid. To determine the occurrence of night-time inversions, I average both surface level and upper air temperature between 2 am and 6 am. If the difference between nightly upper air temperature and surface level temperature is positive, the county experienced a night-time inversion on the particular day. I aggregate the daily data to the time periods of interest (trimesters or in-utero period). I define the instrument as the share of days with a night-time inversion during the time period of interest.

Lastly, I collect reanalysis data on meteorological conditions for the time period 1999-2016

[10] NO2 concentrations are measured at 230 distinct stations over the relevant time period, with a median distance of 21km between county centroids and monitors. The number of active CO monitors is 61. Data can be assigned to a subsample of 122 counties, with a median distance of 32.3 km between county centroids and CO monitors.

[11] Specifically the two data products I use are *ERA5 hourly data on single levels from 1979 to present* and *ERA5 hourly data on pressure levels from 1979 to present*

from the ECMWF, to include as control variables. Monthly average temperature, precipitation and wind speed are available on a 0.1° x 0.1° grid.[12] I aggregate data to counties, using all grid points falling into a county, or in case of small counties without a point on its territory, I assign data from up to 10 closest points within 40km distance using inverse distance weighting. I aggregate the weather data from the monthly level to the time periods of interest.

The pollution and inversion variables for the sample are summarized in the bottom part of Table 2.1. I report concentrations of PM10, NO2 and CO as well as the frequency of inversions for the in-utero period, which is defined as the 270 day or 9-month period ending with but including the month of birth.

2.4. Empirical Strategy

To identify the effect of air pollution exposure on children's socioemotional skills, I rely on an approach based on thermal inversions, following e.g. Arceo et al. (2016); Colmer et al. (2021b); Jans et al. (2018b) and Molina (2021).

The baseline model is given by:

$$NC^{a}_{icym} = \beta PM_{cym} + \gamma' \mathbf{X}_i + \delta' \mathbf{W}_{cym} + \theta_c + \theta_y + \theta_m + \theta_a + u_{icyma}\,, \tag{2.1}$$

where NC^{a}_{icym} denotes a measure of non-cognitive ability of individual i, born in county c in month m of year y. The age group at which the skills are assessed is represented by a. PM_{cym} is local PM10 concentration during the gestational period. The model includes county fixed effects θ_c to account for persistent difference in pollution and skill levels across locations, month-of-birth fixed effects θ_m to control for seasonality in air quality, and year-of-birth fixed effects θ_y to capture changes in cognitive skills over time which affect individuals in all counties equally. θ_a controls for age-specific trends in outcomes. The model further includes individual and family background characteristics $\mathbf{X}_i$ (child gender, age in month and its square, migration background and dummies for parental education levels) and meteorological conditions, $\mathbf{W}_{cym}$. These include third order polynomials in temperature, precipitation and wind speed during the gestational period. Standard errors are clustered at the county level.

For the OLS estimate of β to be consistent, the unobserved determinants of childhood non-cognitive abilities summarized in u_{icyma} must not be correlated with PM10 levels conditional on fixed effects and control variables. This assumption is probably violated, e.g. due to region specific economic shocks affecting both air quality as well as parental income which might be spent on investments into the child's skill development. Secondly, individual pollution

[12]Product: *ERA5-Land monthly averaged data from 1981 to present*

exposure is measured with error which leads to attenuation bias. To address these issues, I rely on the meteorological phenomenon of thermal inversions to extract exogenous variation in air quality. Under normal conditions, air temperature decreases with altitude. Emissions released at the ground level rise and disperse in the air. During an inversion, air temperature increases with altitude, i.e. upper air layers are warmer than ground level air. The warm upper air layer acts like a ceiling that prevents ground level pollution from rising and dispersing. As pollutants are trapped beneath the warm air layer, their surface level concentration increases.

Thermal inversions are a meteorological phenomenon. They exhibit a seasonal pattern with inversions occurring more frequently in the winter as compared to the summer. However, conditional on month-of-birth fixed effects and weather controls, it is as good as random whether the specific combination of meteorological conditions occurs that gives rise to a thermal inversion. Importantly, the frequency of inversions should be plausibly uncorrelated with local business cycles. At pollution levels that were common in Germany during my sample period, inversions usually do not lead to visible smog events or extremely poor air quality, and are thus unlikely to trigger avoidance behavior. Besides, following e.g. Jans et al. (2018b) and Molina (2021), I exclusively consider nighttime inversions, which should be even less likely to induce any behavioral responses.

Specifically, the first stage model is given by:

$$PM_{icym} = \alpha Inv_{cym} + \gamma' \mathbf{X}_i + \delta' \mathbf{W}_{cym} + \theta_c + \theta_y + \theta_m + \theta_a + u_{icyma} \tag{2.2}$$

where Inv_{cym} is the share of days on which a nighttime inversion occurred during the period of interest. In addition to the variables mentioned above, I add one lead and one lag of the instrument to $\mathbf{W}_{cym}$ in both the first and second stage model, to account for autocorrelation in inversion frequency.

The first stage results in Table 2.2 replicate the finding from previous studies that thermal inversions are a strong instrument for PM10, thus satisfying the relevance condition. The effect magnitude in columns 1 and 2 imply that mean PM10 concentration during a 9 month period increases by 1.1 and 1.2 μg/m^3, respectively, in the two analysis samples, for a standard deviation increase in inversion frequency (.07).[13] The associated F-Statistics are large. In columns 3 to 6, I show results when replacing PM10 by either NO2 or CO as outcome. Inversions have a somewhat smaller, but highly significant effect on NO2, which is a common co-pollutant of PM10. For CO, the effect is also positive, but less significant, such that the F-Statistic is clearly below the common threshold for a sufficiently strong instrument. In sum, thermal inversions

[13]The coefficients in the table depict the effect for moving from no inversions at all to having an inversion each night during the period of interest. In the data, however, none of these two extreme cases occurs, and the actual variation in inversion frequency is substantially smaller.

increase the concentration of all three pollutants considered in the table, and with only one instrument, I cannot cleanly disentangle their distinct effects on noncognitive skills. However, the first stage results suggest that PM10 is most likely to drive any results in the second stage, because (i) unlike CO and fine particles, NO2 is not commonly considered as a major risk factor in the medical literature, and (ii) the first stage effect for CO is weak.

Table 2.2.: First Stage Results

	PM10	PM10	NO2	NO2	CO	CO
Share Inversions	15.86***	17.82***	10.85***	9.87***	122.12	187.93**
in utero	(1.611)	(1.985)	(1.223)	(1.535)	(87.679)	(90.747)
Observations	7,645	4,911	8,880	6,099	5,494	3,679
F-Statistic	98.68	81.36	73.63	41.47	1.91	3.91
Sample	I	II	I	II	I	II

Note: The table depicts estimates of the effects of inversion frequency during the nine-month in-utero period on pollution concentration during that period from model 2.2. All outcomes are measured in $\mu g/m^3$. Regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, weather controls and individual background characteristics as described in the text. Standard errors clustered at the county level are reported in parentheses. Significance levels: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The key identifying assumption underlying the 2SLS and reduced form estimations is that, conditional on the included covariates, the frequency of thermal inversions during the in-utero period affects noncognitive skills in childhood only through their effect on air pollution. This assumption will be violated if families of children exposed to more inversion periods during early life differ systematically from families of children exposed to fewer inversions along unobservable characteristics.

As a test of the identifying assumption, I check whether inversion frequency is systematically correlated with observed predetermined family background variables. The results of these falsification checks are displayed in Table 2.3. The estimated coefficients reflect the change in outcomes for a one standard deviation increase in inversion frequency (0.07) and reveal that family characteristics show no consistent pattern with respect to inversion frequency: For instance, fathers of children exposed to more inversions while in-utero tend to be slightly less educated, but at the same time, on average, slightly fewer of these children have a migration history. Importantly, none of the tested variables (maternal and paternal education, maternal age at birth and migration background) are significantly correlated with the instrument and the point estimates are generally small in magnitude.

Table 2.3.: Falsification Tests

	Maternal Age at birth	Migration History	Tertiary degree (Mother)	Less than High School (Mother)	Tertiary degree (Father)	Less than High School (Father)
Panel A: Sample I						
Share Inversions	-.1691	-.0054	-.0177	.0080	-.0150	.0137
in utero	(.1570)	(.0117)	(.0113)	(.0098)	(.0163)	(.0128)
Observations	9,200	9,440	9,201	9,201	6,284	6,284
Panel B: Sample II						
Share Inversions	-.1861	-.0076	-.0042	-.0014	-.0064	.0130
in utero	(.1963)	(.0144)	(.0136)	(.0116)	(.0184)	(.0151)
Observations	6,326	6,548	6,380	6,380	4,157	4,157

Note: The table depicts coefficients from OLS regressions of family characteristics on the share of inversions during the child's in-utero period. Maternal age is measured in years. The other outcomes are dummy variables. Regressions control for cubic functions of precipitation, temperature and wind speed. Standard errors clustered at the county level are reported in parentheses. Significance levels: *p<0.1; **p<0.05; ***p<0.01

2.5. Results

This sections starts by presenting the 2SLS and reduced form results for the effects of in-utero exposure to air pollution on the Big Five personality traits. The main finding is that worse air quality reduces emotional stability. Thereafter, I explore this result in terms of heterogeneity and critical phases of exposure, and test its robustness to changes in sample construction, the set of included covariates and the choice of the instrument.

2.5.1. Main results

Panel A of Table 2.4 displays 2SLS estimates of the effect of in-utero exposure to PM10 on the Big Five personality traits. A large and significant impact is found for neuroticism which increases by 7.1% of a standard deviation for an increase in mean PM10 concentration by one $\mu g/m^3$. This implies that individuals exposed to higher levels of air pollution during the prenatal period are less emotionally stable in childhood, i.e. more fearful and less self-confident. The second largest point estimate emerges for openness. As mentioned earlier, openness as measured in the SOEP most likely partially reflects cognitive skills. Thus, the negative estimate is in line with existing results on the negative impact of in-utero exposure to air pollution on later life cognitive ability. The fact that the negative effect is not statistically significant is unsurprising given that the questions asked to assess openness are only crude measures of cognitive ability as compared to detailed tests used in previous studies. The remaining three traits are not affected by gestational particulate matter levels. The point estimates are close to zero (all below 1.7% of a standard deviation), and not statistically significant.

These results are confirmed by the estimates from the reduced form models, which are presented in panel B. As mentioned above, the reduced form can be estimated on larger samples. Coefficients are multiplied by 0.07 to reflect the estimated effects for a one standard deviation increase in inversion frequency. Based on the first stage results this corresponds to an increase in PM10 by a bit more than one μg/m^3. In line with the 2SLS estimates, the only significant result emerges for neuroticism, which rises by 6.7% of a standard deviation in response to the increase in in-utero inversion frequency. This is very similar in terms of magnitude to the 2SLS estimate, implying that the relationship holds in counties with and without PM10 monitors. Regarding the other four traits, the reduced form confirms the absence of any effects. Again, apart from the negative but insignificant coefficient for openness, the point estimates are small in magnitude. A potential explanation for these null effects could be that these traits are measured on a different sample than neuroticism, including younger children. Possibly, personality differences in 2-3-year-olds are not yet pronounced enough or the measures used in the SOEP for this age group are too crude to uncover the impact of prenatal pollution exposure. Therefore, in Table 2.A.2, I repeat the estimation for these four outcomes on *sample II*, i.e. the sample used for the analysis of neuroticism, including only children aged 5 or older. The results are similar, again showing no significant effect on any of the four traits. The pattern is also unchanged, showing the largest negative point estimates for openness and relatively small estimates for the other three outcomes.

2.5.2. Additional Results

Having established that in-utero exposure to particulate matter increases neuroticism in childhood, this paragraph further examines this effect in terms of critical windows of exposure and effect heterogeneity.

Effect Heterogeneity. I analyze effect heterogeneity with respect to child and family characteristics, namely child gender, age at which neuroticism is assessed (5-6 vs. 9-10 years), and current household income (above vs. below median). I do not find evidence for heterogeneity in the effect of inversions on neuroticism along any of the three dimensions. Results are

Table 2.4.: Impact of prenatal PM10 exposure on the Big Five

	Openness	Consc.'ness	Extraversion	Agreeableness	Neuroticism
Panel A: 2SLS					
PM10	-.0361	.005	.0033	.0163	.0714**
in utero	(.0288)	(.0279)	(.0258)	(.0261)	.0314
Observations	7,655	7,645	7,666	7,614	4,911
Counties	163	163	164	162	122
1st Stage F-Stat.	98.	96.9	98.7	95.1	80.5
Panel B: Reduced Form					
Inversions	-.0423	-.0182	-.0034	.0231	.0667**
in utero	(.0279)	(.0264)	(.0261)	(.0265)	(.0323)
Observations	9,454	9,440	9,445	9,423	6,548
Counties	192	192	192	192	155

Note: Panel A displays 2SLS-estimates of the effect of a 1 μg/m^3 increase in PM10 concentration during the in-utero period on Big Five personality traits. PM10 is instrumented by inversion frequency. Panel B shows the OLS estimates of the effect of an increase in the share of days with a nighttime inversion during the in-utero period by one standard deviation (+7%) on the outcomes. Outcomes are standardized within age groups. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

illustrated in Appendix Figure 2.B.4. In all three cases, the coefficient on the interaction term is close to zero and insignificant, whereas the main effect remains statistically significant and of similar magnitude as in the baseline model. In summary, the effects of prenatal air pollution exposure emerge already by school start age, affect both genders equally and are not dampened by parental resources.

Secondly, I examine heterogeneity in effect magnitude along the distribution of neuroticism. I construct two indicator variables for the value of neuroticism falling below the first quartile and above the third quartile, respectively. Figure 2.3 shows 2SLS and reduced form results. While all point estimates have the expected sign, i.e. on average air pollution exposure reduces the probability to fall below the lower quartile and increases the probability to fall above the upper quartile, only the latter effect is statistically significant and also larger in magnitude (-1.5 pp vs +3 pp). This result is noteworthy as it implies that prenatal air pollution might affect child mental health: In modern personality psychology, mental disorders are conceptualized as extreme realizations of the Big Five traits. High values of neuroticism show particularly robust and consistent correlations with a range of mental health issues, e.g. depression and anxiety disorders. (Almlund et al., 2011; Widiger et al., 2017) Hence, the fact that the main result is mostly driven by increases in neuroticism in the upper part of the distribution suggests that prenatal particulate matter exposure might not just reduce an important noncognitive skill within the range of 'normal' variations in personality, but could even give rise to mental health issues. It should be stressed that this is only suggestive, and with the available data, I am unable to explore this in more depth.[14]

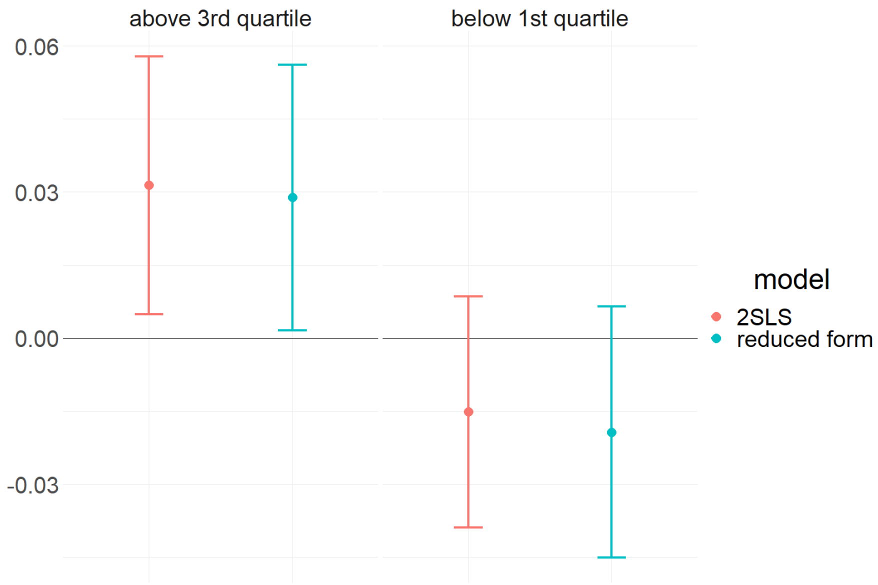


Figure 2.3.: Effect Heterogeneity along the distribution of Neuroticism

Note: Red colored dots depict 2SLS estimates of the effect of a 1 μg/m^3 increase in PM10 concentration during the in-utero period on indicator variables taking the value one if age standardized neuroticism falls above the 75th percentile of the sample distribution (left) or below the 25th percentile of the sample distribution (right), respectively. Blue colored dots depict OLS estimates of the effect of a one standard deviation increase in the share of days with a nighttime inversion during the in-utero period on the same outcomes. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Bars represent 95%-confidence intervals, based on standard errors clustered at the county level.

Trimester level effects. To assess critical windows of exposure, I split the in-utero period into three trimesters, and run the reduced form regression including variables measuring inversion frequency separately for each trimester. Besides, to analyze whether postnatal air pollution exposure also generates adverse long-run effects on neuroticism, the regression includes inversions during the first nine month after birth, broken down into three-month periods. Symmetrically, I include inversion frequency during three *placebo trimesters*, i.e. three-month periods preceding conception. Assuming that air pollution affects childhood neuroticism solely through physiological channels, the frequency of thermal inversions before conception should not have a significant impact. Regressions include the same fixed effects, indi-

[14]The SOEP questionnaires do not include instruments measuring child mental health, and formal diagnosis of mental disorders is a very rare outcome that cannot sensibly be analyzed given the sample size.

vidual and family characteristics as before and trimester-specific, quadratic weather controls. Table 2.5 presents the results. Significant positive effects are found for inversion frequency during the second and third trimester of pregnancy. Inversions during the first trimester and the first nine month after birth have no significant impact on neuroticism. This result resembles the findings on the effect of air pollution on performance in school achievement tests in Bharadwaj et al. (2017) (effects mainly in the third trimester) and cognitive ability in Molina (2021) (significant effects only in the second trimester). Reassuringly, the coefficients on the three placebo trimesters are insignificant and also of smaller magnitude than the effects of in-utero exposure.

Table 2.5.: Neuroticism - Effects by Trimester

	Neuroticism
Inversions	-.0050
9-7 months before conception	(.0165)
Inversions	.0227
6-4 months before conception	(.0167)
Inversions	.0179
3-1 months before conception	(.0163)
Inversions	-.0079
1st trimester	(.0187)
Inversions	.0327**
2nd trimester	(.0164)
Inversions	.0384**
3rd trimester	(.0169)
Inversions	-.0231
1-3 months post birth	(.0176)
Inversions	.0171
4-6 months post birth	(.0163)
Inversions	-.0220
7-9 months post birth	(.0169)
Observations	6,548

Note: The table shows OLS estimates of the effect of an increase in the share of days with a nighttime inversion during the in-utero period by one standard deviation (+7%) on Neuroticism (age-standardized). Regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, as well as covariates for parental education, child gender, migration background, age in months and its square, plus trimester-specific, quadratic functions of temperature, wind speed and precipitation. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Postnatal exposure. The reduced form results just presented indicate that neuroticism is not affected by postnatal exposure to air pollution. To corroborate this finding, I analyse the effects of air pollution during the first nine month after birth in the main 2SLS and reduced form models, for neuroticism as well as the other noncognitive skills. Since early life is also a critical period of development, I run these regressions to make sure not to miss any relevant effects. 2SLS and reduced form models are defined as before, with the only difference that weather controls are included for both the pre- and postnatal period to control for any possible long-run effects of early life weather conditions. I find no effects of postnatal PM10 exposure on any of the five outcomes. As mentioned above, the finding that noncognitive ability, in particular neuroticism, is only affected by prenatal pollution exposure is in line with results for cognitive ability. The result is also plausible given the fact that both types of skills are governed by brain structure and function which develops most rapidly in the gestational period. However, adverse long-run effects on educational and labor market outcomes where found for exposure during both the in-utero period and the first year of life (Isen et al., 2017b; Voorheis, 2017). This implies that other components of human capital, e.g. physical health, might also mediate a part of the adverse long-run effects. This is in line with results by Klauber et al. (2021) who follow children from birth to age five and find that in-utero exposure to particulate matter causes subtle but persistent damage to respiratory health.

Robustness Checks. The main finding from the preceding paragraphs is that in-utero exposure to particulate matter increases neuroticism at ages 5-10. The similarity of the 2SLS and reduced form results as well as the absence of an effect during the placebo trimesters provide a first confirmation that the finding is not driven by correlated unobservables. However, given the modest sample size and the large number of hypotheses tested, I subject this finding to a number of additional robustness checks.

First, I test sensivity to changes in sample construction. Results are reported in Appendix Table 2.A.4. Panel A shows 2SLS estimates, while panel B presents reduced form estimates. Column 1 replicates the baseline result. In columns 2 and 3, I vary the minimum number of observations for a county to be included in the sample to either 15 (column 2) or 25 (column 3) instead of 20. While the number of included individuals and counties changes notably across the three columns, the point estimates are positive and significant in all cases, with their size varying slightly between 6.65% and 8.88%. The fact that the estimated coefficient gets larger and more significant for the larger cut-off value is plausible as it is easier to precisely estimate the effect of pollution exposure if the within-county comparison group is larger. In columns 4 and 5, I use all available observations in the SOEP, instead of keeping only one observation per individual. If a child is observed at both 5-6 and 9-10 years of age, both observations are included. Column 4 reports results from an unweighted regression. In column 5, I weight observations by the inverse of the number of observations per individual. All estimates are similar in quantitative terms to the baseline result, implying that the finding is not driven by the specific choice of the analysis sample.

Second, I analyze the robustness of the results to the set of included covariates. Results are displayed in Appendix Table 2.A.5. The baseline result is replicated in column 1. Column 2 shows estimates from a specification that includes a more comprehensive set of background controls. Specifically, I add an indicator for a single parent household, dummies for maternal age at birth (in 5 year steps), and birth order dummies. Columns 4 and 5 depict results from specifications including more and less leads and lags of the instrument, respectively. In all three model versions, the estimated effects are similar in both magnitude and significance to the baseline finding. This holds for both the 2SLS and reduced form results.

Next, I test whether the result could be just a statistical artifact, arising because certain mothers have difficulties in assessing their children's non-cognitive skills. This concern arises from the fact that intermediate values are chosen 'too often' when mothers assess their children's personality traits, and more so by less-educated mothers (see Section 2.3.1). Column 2 of Table 2.A.6 presents results from estimation after dropping observations from the sample where mothers picked the intermediate value of 5 for all items underlying conscientiousness and neuroticism, the two outcomes that showed most excess mass for the middle value (see Figure 2.1). The estimated effect remains statistically significant and drops by less than .2 percentage points relative to the baseline result in column 1 in both the 2SLS- and reduced form specifications. Since this approach only drops a small number of observations from the sample, I also employ an alternative strategy: I compute the variance in responses across the ten questions underlying the Big Five personality traits in the SOEP, after demeaning all items in the full sample. I then drop the 5% (10%) of the sample with the lowest variance. This reflects the idea that mothers who are good at assessing their children's personality will deviate more from the population mean. This approach is partly based on Falk et al. (2021).[15] The results are depicted in columns 3 and 4. The effect of exposure to air pollution during the in-utero period remains statistically significant in both cases. Both in the reduced form and the 2SLS specification, it increases slightly in magnitude relative to the baseline model. Overall, responses by

[15]In the paper, Falk et al. (2021) build a choice model of survey response behavior where respondents have imperfect self-knowledge. They develop an estimator of self-knowledge and show that regressions using self-reported risk-attitudes or non-cognitive skills on either the left or right side yield larger estimated coefficients and R^2 in the subsample of individuals with higher self-knowledge. In my application, mothers answer questions about their children's personality. Hence, the term self-knowledge is not suitable here, but the general issue is the same: mothers differ in their capability to memorize and recall relevant information about child behavior, or in their knowledge about 'normal' child behavior. Mothers with the lowest levels of this type of knowledge will likely opt for intermediate values. As the proposed estimator of self-knowledge requires panel data, which is not available for all individuals in my sample, I can only estimate the between-variance, the numerator of the estimator for self-knowledge by Falk et al. (2021) which rises monotonically in self-knowledge in their model.

mothers who have difficulties in evaluating their child's noncognitive ability seem not to bias the results.

In a final robustness check, I re-run the 2SLS and reduced form models with a different instrument, the inverse of the planetary boundary layer height (PBLH). This instrument was recently used by Godzinski and Castillo (2021). The planetary boundary layer (PBL) is the lowest part of the troposhere, i.e. the air layer directly above the surface. Pollution emitted at the ground level disperses within this air layer. The higher the PBL, the larger the volume of air in which emissions can dissipate, leading to lower ground level concentration of air pollutants. Hence, for lower values of the PBLH, or higher values of the inverse of the PBLH, pollution concentration near the surface increases. This implies that PBLH exploits the same physical mechanism to extract exogenous variation in air pollution as thermal inversions and thus also affects multiple air pollutants. However, while the main instrument only measures whether on any given day a thermal inversion either occurs or does not occur, the inverse of the PBLH is a continuous variable accounting for the strength of the meteorological phenomenon. To construct the alternative instrument, I collect data on average monthly PBLH (in 1000 meters) from the ECMWF, aggregate this to counties and the time period of interest as described in Section 2.3.2, and then take the inverse of the resulting variable.[16] The correlation of inversion frequency and inverse PBLH is 0.266 in *sample II*. Table 2.A.7 shows 2SLS and reduced form results based on the inverse PBLH. Estimated coefficients are very similar in magnitude to the baseline results. The estimates are only significant at the 10% level, which likely follows from the fact that the inverse PBLH is not as strong an instrument as the frequency of nighttime inversions (F-Statistic of 18.5). Overall, I view this as a confirmation of the main results, indicating that the effect on neuroticism is not driven by unobservable variables correlated with inversion frequency.

Discussion. The main result of the analysis is that neuroticism increases by approximately 7% of a standard deviation for a one unit increase in prenatal PM10 concentration. Assessing whether it is plausible from a medical perspective to find an effect of in-utero exposure to particulate matter on neuroticism but not the other dimensions of non-cognitive ability, and discussing potential channels for this effect, is difficult given that the research on neurobiological origins of personality traits is still in a rather early stage (Allen and DeYoung, 2017). However, the existing results suggest maternal cortisol levels as a potential pathway between particulate matter exposure during pregnancy and the child's level of neuroticism. Li et al.

[16] The mean value of the inverse PBLH in *sample II* is 1.76, with a standard deviation of 0.2. The effect of a one unit increase in the inverse PBLH on PM10 concentration during the in-utero period is 4.48. Hence a one standard deviation increase in the IV raises PM10 by 0.9 $\mu g/m^3$.

(2017) find in a randomized experiment with college students in Shanghai that acute exposure to PM2.5 causes an increase in cortisol and other stress hormones, pointing to activation of the hypothalamus-pituitary-adrenal axis. Cortisol and activity in the hypothalamus-pituitary-adrenal axis in turn show relatively robust associations with neuroticism. (Allen and DeYoung, 2017) Cortisol as potential link between air pollution and neuroticism is also in line with the result by Persson and Rossin-Slater (2018) that maternal stress during pregnancy has a causal impact on the mental health of children, including adult anxiety and depression.

To assess the magnitude of the estimated effect, a comparison to related studies examining the impact of gestational pollution exposure on later life cognitive ability is helpful. Results by Molina (2021) imply that a one $\mu g/m^3$ increase in PM10 concentration during the second trimester reduces cognitive ability in adulthood, measured by the Raven's test, by 2.4% of a standard deviation.[17] Sanders (2012) finds that among cohorts born during the early 1980s in the US, a 10 $\mu g/m^3$ increase in total suspended particulate (TSP, which includes both PM10 and larger, less harmful particles) exposure during the students' year of birth reduces their high school math test scores by 6% of a standard deviation. Given an average PM10-to-TSP ratio across North America of approximately 0.5 (Cicero-Fernandez et al., 1993; Van der Meulen et al., 1987) , this implies that a 1 $\mu g/m^3$ increase in PM10 concentration would reduce test scores by approximately 1.2% of a standard deviation. To account for the fact that the mean age of survey respondents studied in Molina (2021) is 17, and test takers considered in Sanders (2012) are aged 15-18, whereas children in my sample are assessed between age 5 and 10, I rescale my main 2SLS estimate (Table 2.4, column 5) with the correlation between neuroticism in childhood and at ages 16-17 (Figure 2.2). This implies that a 1 $\mu g/m^3$ increase in prenatal PM10 concentration raises neuroticism at that age range by 7% x 0.137 = 1% of a standard deviation. While my analysis is conducted in a very different setting with substantially lower baseline pollution, the result is of the same order of magnitude as the effects found on cognitive skills. Hence, non-cognitive ability is a plausible additional mechanism driving a part of the adverse effects of in-utero exposure to air pollution on labor market outcomes.

To further assess the relevance of this channel, I conduct a back-of-the-envelope calculation to approximate the earnings impact of poor air quality via the increase in neuroticism. Using survey data and a sibling fixed effects approach, Fletcher (2013) finds that a one standard deviation increase in neuroticism measured in young adulthood reduces annual earnings by 5-6%. As mentioned above, my results imply that a 1 $\mu g/m^3$ increase in gestational PM10 exposure raises neuroticism during young adulthood by 1% of a standard deviation. Combining these

[17] This number is derived by combining quasi first stage- and reduced form results in Molina (2021). Hence, it should be interpreted with caution. However, for the purpose of assessing whether my results are of a comparable order of magnitude, I think computing this estimate is justified and helpful. PM10 levels during the first and last trimester of pregnancy have no effect on the Raven's test score.

results indicates that a one standard deviation (5 μg/m^3) increase in prenatal PM10 concentration reduces annual earnings by .24%-.29% through its adverse effect on emotional stability. Isen et al. (2017b) exploit the US clean air act and find that a reduction in early life exposure to TSP by 10 μg/m^3 increases adult earnings by 1%. While a comparison of results derived from different settings and time periods of course has limitations, it at least suggests that the noncognitive ability channel is of relevant size. Given that my estimates are likely attenuated due to some missing information on the place of birth, the channel could account for roughly a third of the full effect.

2.6. Conclusion

This paper provides causal evidence on the effect of in-utero exposure to air pollution on noncognitive ability in childhood. Using the meteorological phenomenon of thermal inversions to address the endogeneity in exposure to particulate matter, I find that exposure to PM10 during the prenatal period reduces emotional stability at age 5-10 in a sample of children born in Germany since the year 2000. In terms of magnitude, an increase in PM10 concentration by 1 μg/m^3 raises neuroticism by 7% of a standard deviation. The result proves robust to changes in the model specification, analysis sample and instrument. Back of the envelope computations imply that an increase in PM10 by 5 μg/m^3, i.e. one standard deviation, reduces adult earnings by .24%-.29% just through its impact on neuroticism.

The finding is important in light of recent evidence showing that the labour market returns to noncognitive skills have increased, relative to the returns to cognitive skills (Edin et al., 2017; Deming, 2017). This suggests that the magnitude of this channel might become even larger over time.

The results are obtained from a setting with relatively low baseline pollution levels, especially in comparison to other studies on long-run effects of in-utero exposure. Hence even at air quality levels below current limit values in developed countries, adverse long-run effects on skills, and thus most likely also later life labor market outcomes, arise.

The finding has also important policy implications in light of the fact that noncognitive ability is malleable during childhood and adolescence. Investments targeted at emotional stability, e.g. mentoring programs, might be a feasible strategy to alleviate the negative impacts of early life exposure to air pollution on educational attainment and adult earnings. Individuals born in places during periods of poor air quality, e.g. from natural sources such as wild fires, could at least in part be compensated for these bad starting conditions by way of targeted interventions. This result might guide policy-makers in allocating scarce resources for programs fostering skill development.

Finding that the rise in neuroticism is mainly driven by increases at the upper end of the distribution suggests that in-utero exposure to air pollution might not only reduce emotional stability within the range of 'normal' variations in personality, but also induce mental health problems. This suggestive relationship between in-utero exposure to air pollution and later life mental health should be explored more in future research. Another interesting avenues for future work would be to replicate this analysis in other settings, e.g. a developing country, to investigate the external validity of the result, or using data on adults to examine how the effect develops over the life-cycle. Lastly, one caveat of this analysis is that the instrument does not allow to disentangle the effects of different pollutants, most importantly CO and PM10. Repeating the analysis with a different instrument that overcomes this issue, e.g. changes in relevant policies, can generate important insights for policy-makers deciding about regulation of air pollutants.

Appendix to Chapter 2

2.A. Additional Tables

Table 2.A.1.: Big Five in the SOEP Mother-and-child-questionnaires

How would you rank your child in comparison to other children of the same age? My child is ...		
2-3 years	O	quick at learning new things - needs more time
	C	focused - easily distracted
	E	shy - outgoing
	A	obstinate - obedient
	N	-
5-6 years 9 -10 years	O	understands quickly – needs more time
		not that interested – hungry for knowledge
	C	focused - easily distracted
		tidy – untidy
	E	talkative – quiet
		withdrawn – sociable
	A	obstinate - compliant
		good-natured – irritable
	N	self-confident – insecure
		fearful – fearless

Note: Questions on child Big Five Personality Traits asked in the SOEP Mother-and-child-questionnaires.

Table 2.A.2.: Impact of prenatal PM10 exposure on the Big Five: Outcomes on Sample II

	Openness	Conscien.'ness	Extraversion	Agreeableness
Panel A: 2SLS				
PM10	-.0398	.0179	-.0228	.0337
in utero	(.0274)	(.0310)	(.0280)	(.0315)
Observations	4,887	4,902	4,879	4,853
Counties	121	122	121	120
First Stage F-Stat	79.6	80.4	80.5	78.0
Panel B: Reduced Form				
Inversions	-.0458	-.0257	-.0171	.0411
in utero	(.0291)	(.0316)	(.0314)	(.0355)
Observations	6,541	6,534	6,532	6,517
Counties	155	155	155	155

Note: Panel A displays 2SLS-estimates of a 1 $\mu g/m^3$ increase in PM10 concentration during the in-utero period on Big Five personality traits, estimated on *sample II*, i.e. children assessed at ages 5-6 or 9-10. PM10 is instrumented by inversion frequency Panel B shows the OLS estimates of the effect of an increase in share of days with a nighttime inversion during the in-utero period by one standard deviation. Outcomes are age-standardized. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age and age squared, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Standard errors clustered at the county level are reported in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 2.A.3.: Impact of postnatal PM10 exposure on the Big Five

	Openness	Consc.'ness	Extraversion	Agreeableness	Neuroticism
Panel A: 2SLS					
PM10	.0256	.0017	.0096	.0202	-.0251
post birth	(.0224)	(.0242)	(.0219)	(.0263)	(.0283)
Observations	8,081	8,070	8,074	8,016	5,354
Counties	169	169	169	167	131
1st Stage F-Stat.	108.2	107.7	108.4	106.8	87.6
Panel B: Reduced Form					
Inversion Frequency	.0397	-.00002	.0104	.0316	-.0181
post birth	(.0263)	(.0241)	(.0240)	(.0286)	(.0346)
Observations	9,454	9,440	9,445	9,423	6,548
Counties	192	192	192	192	155

Note: Panel A displays 2SLS-estimates of the effect of a 1 μg/m^3 increase in PM10 concentration during the nine-month period following the month of birth on Big Five personality traits. PM10 is instrumented by inversion frequency. Panel B shows the OLS estimates of the effect of an increase in the share of days with a nighttime inversion during the postnatal period by one standard deviation (+7%) on the outcomes. Outcomes are age-standardized. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age and age squared, cubic functions of temperature, wind speed and precipitation during both the postnatal and prenatal period, plus one lead and one lag of the instrument. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.4.: Robustness - Sample Construction

	Dependent Variable: Neuroticism				
	(1)	(2)	(3)	(4)	(5)
Panel A					
PM10	.0714**	.0665**	.0888***	.0823***	.0725**
in utero	(.0314)	(.0281)	(.0343)	(.0289)	(.0294)
Minimum obs. per county	20	15	25	20	20
Obs. per individual	1	1	1	all	all
Weights	x	x	x	x	✓
Observations	4,911	5,600	4,293	6,572	6,572
Counties	122	163	94	122	122
1st Stage F-Stat.	80.7	106.9	63.2	89.0	81.0
Panel B					
Inversions	.0667**	.0697**	.0887***	.0768**	.0632**
in utero	(.0323)	(.0307)	(.0335)	(.0300)	(.0304)
Minimum obs. per county	20	15	25	20	20
Obs. per individual	1	1	1	all	all
Weights	x	x	x	x	✓
Observations	6,548	7,230	5,747	8,497	8,497
Counties	155	196	118	155	155

Note: The table displays 2SLS-estimates of the effect of a 1 μg/m^3 increase in in-utero PM10 exposure on Neuroticism (Panel A) and reduced form results (Panel B). The outcome is standardized within age groups. Column 1 replicates the baseline results. Columns 2 and 3 vary the minimum number of individuals required within a county. In columns 4 and 5 multiple observations for the same individual are included, if available. In column 5 observations are weighted by the inverse of the number of observations per individual. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.5.: Robustness - Alternative Specifications

	Dependent Variable: Neuroticism			
	(1)	(2)	(3)	(4)
Panel A				
PM10	.0714**	.0666**	.0726**	.0771**
in utero	(.0314)	(.0313)	(.0330)	(.0357)
Specification	*baseline*	*extended controls*	*no leads & lags*	*2 leads & lags*
Observations	4,911	4,882	4,911	4,911
Counties	122	122	122	122
1st Stage F-Stat.	80.7	81.3	96.2	43.5
Panel B				
Inversions	.0667**	.0640*	.0674**	.0841**
in utero	(.0323)	(.0324)	(.0326)	(.0352)
Specification	*baseline*	*extended controls*	*no leads & lags*	*2 leads & lags*
Observations	6,548	6,496	6,548	6,415
Counties	155	154	155	152

Note: The table displays 2SLS-estimates of the effect of a 1 μg/m^3 increase in in-utero PM10 exposure on Neuroticism (Panel A) and reduced form results (Panel B). The outcome is standardized within age groups. Column 1 replicates the baseline results. Column 2 adds dummies for single parent household, birth order and maternal age. Columns 3 and 4 vary the number of leads and lags of the instrument included in the regression. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.6.: Robustness - Uninformative Anwers

	Dependent Variable: Neuroticism			
	(1)	(2)	(3)	(4)
Panel A				
PM10	.0714**	.0698**	.0717**	.0811**
in utero	(.0314)	(.0309)	(.0336)	(.0348)
Sample	*baseline*	*drop answers with many "fives"*	*drop bottom 5% (variance in answers)*	*drop bottom 10% (variance in answers)*
Observations	4,911	4,882	4,453	4,125
Counties	122	122	113	108
1st Stage F-Stat.	80.7	83.5	75.4	68.9
Panel B				
Inversions	.0667**	.0657**	.0696**	.0687*
in utero	(.0323)	(.0325)	(.0351)	(.0383)
Sample	*baseline*	*drop answers with many "fives"*	*drop bottom 5% (variance in answers)*	*drop bottom 10% (variance in answers)*
Observations	6,548	6,512	5,943	5,453
Counties	155	155	144	134

Note: The table displays 2SLS-estimates of the effect of a 1 $\mu g/m^3$ increase in in-utero PM10 exposure on Neuroticism (Panel A) and reduced form results (Panel B). The outcome is standardized within age groups. Column 1 replicates the baseline results. In column 2 observations for which all items underlying Neuroticism and Conscientiousness have a value of 5 are dropped. In columns 3 and 4 the 5% or 10% of observations with the lowest variance across items underlying the Big Five in the SOEP are dropped. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation, plus one lead and lag of the instrument. Standard errors clustered at the county level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.7.: Robustness - Alternative IV

	Dependent Variable: Neuroticism	
	(1)	(2)
PM10	.0947*	
in utero	(.0552)	
Inverse PBLH		.0712*
in utero		(.0403)
Observations	4,911	6,548
Counties	122	122
1st Stage F-Stat.	18.5	
1st Stage Effect ($\hat{\alpha}$)	4.48	
Inverse PBLH	Mean	St Dev.
	1.76	.20

Note: Column 1 displays results from 2SLS estimation of the effect of a 1 μg/m^3 increase in in-utero PM10 exposure on Neuroticism, using the inverse planetary boundary layer height as an instrumental variable. Column 2 reports results from the reduced form model, multiplied by .2 to reflect the effect of a one standard deviation increase in the inverse PBLH. The outcome is age-standardized Neuroticism. Regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age in months and its square, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Standard errors clustered at the county level are reported in parentheses. The bottom part shows summary statistics for the instrumental variable. *p<0.1; **p<0.05; ***p<0.01

2.B. Additional Figures

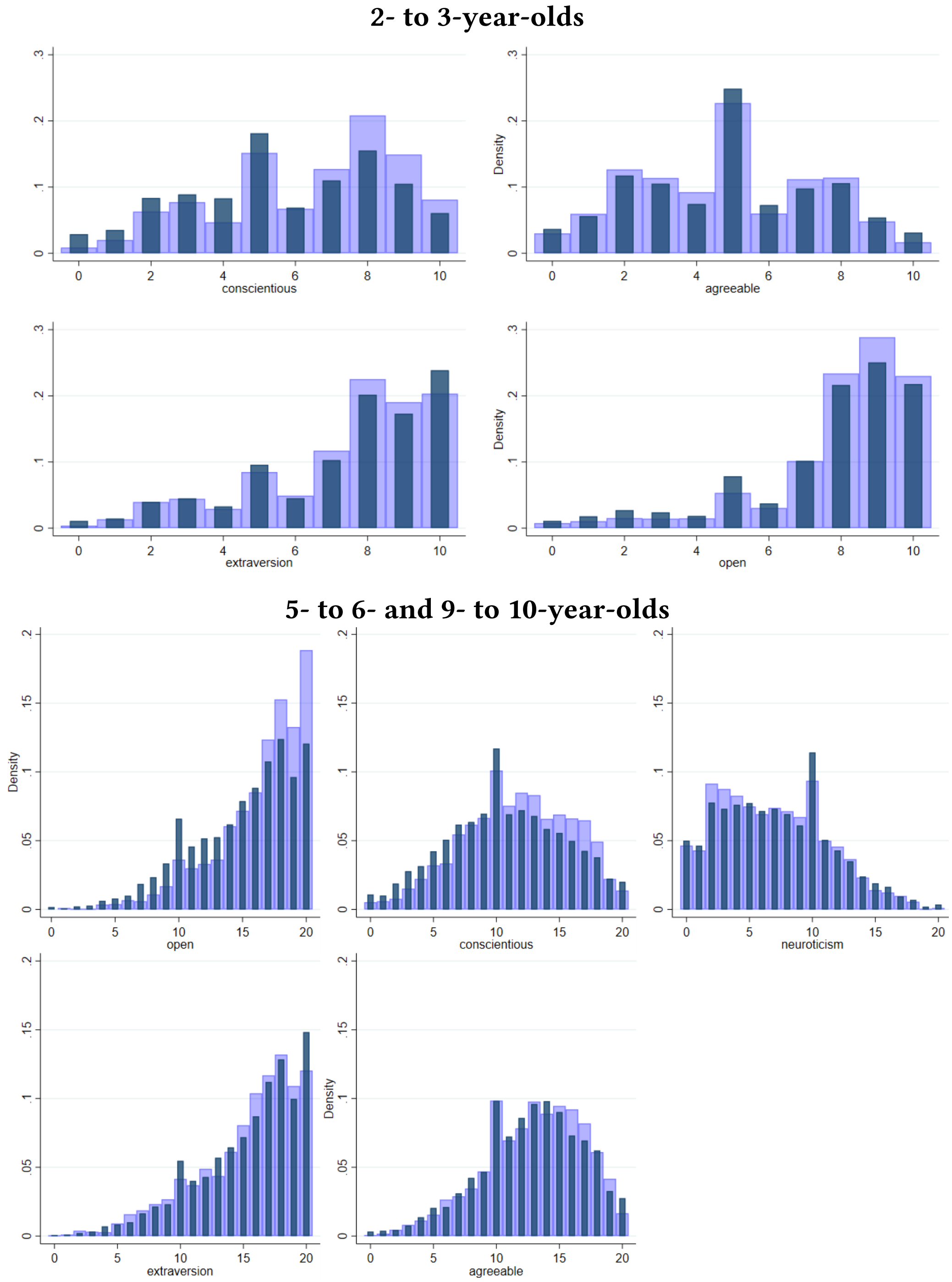


Figure 2.B.1.: Distribution of Big Five personality traits by maternal education.

Note: The distributions are based on all available observations in the SOEP. Light blue [Dark blue] bars represent the distribution among children whose mother holds a [holds no] tertiary degree. Top: 2,185 answers by highly-educated mothers, 5,803 by less-educated mothers. Bottom 2,850 answers by highly-educated mothers, 8,518 by less-educated mothers.

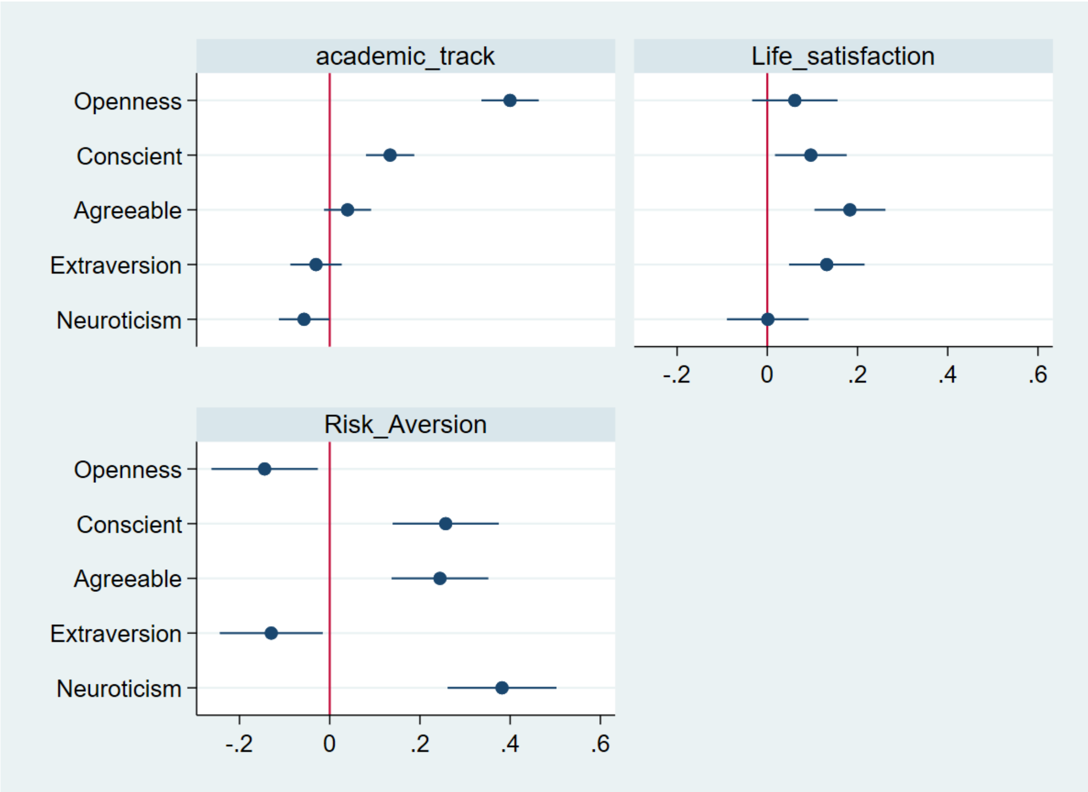


Figure 2.B.2.: Non-cognitive abilities and child outcomes

Note: The figures show partial correlations between mother-assessed Big Five and other child outcomes measured in the SOEP, after controlling for parental education, a single-parent household dummy, child gender and migration background. Upper left plot: Correlations between standardized mother-reported Big Five when the child is aged 5-6 years and the mother-assessed probability that the child will graduate from the academic track of the German school system, measured on an 7-point scale when the child is 7 to 8 years old. Sample size: 3,842. Upper right and bottom plots depict correlations between standardized mother-reported Big Five when the child is aged 9-10 years and child-reported life satisfaction and risk aversion at age 11-12, respectively. Both are measured on an 11-point scale. Sample Sizes: 2,210 and 2,190, respectively. 95%-Confidence Intervalls are based on robust standard errors. Based on data from SOEP, version 35.

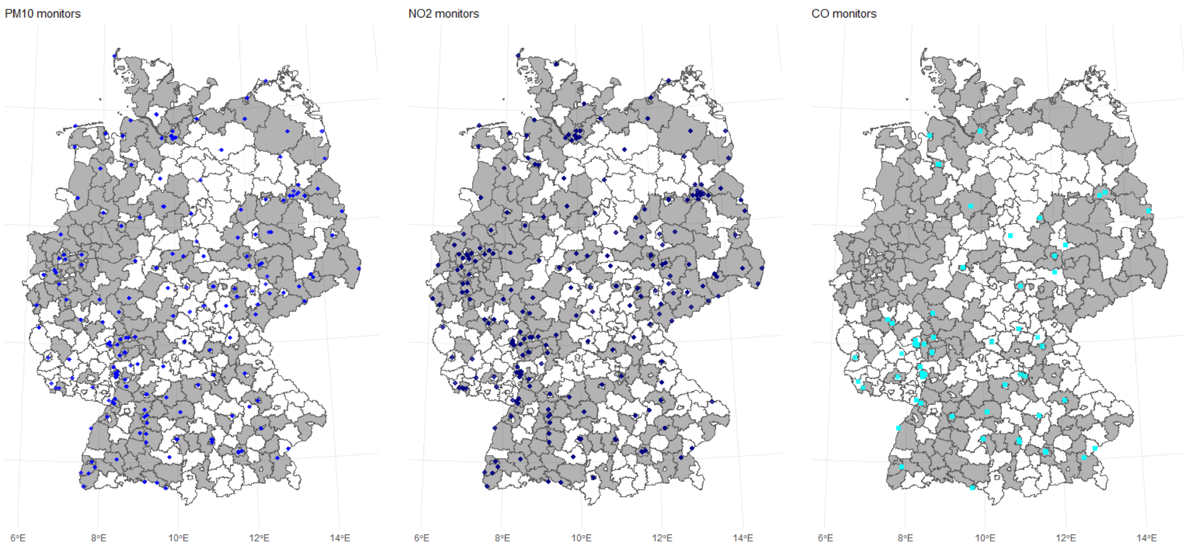


Figure 2.B.3.: Sample Counties and pollution monitors

Note: Grey shaded counties are those included in *Sample I*. Based on data from UBA and SOEP, version 35. Dots represent pollution monitors which were active throughout the sample period and are used in the analysis. Blue dots on the left map represent PM10 monitors, dark blue dots in the middle map represent NO2 monitors, light blue dots on the right map represent CO monitors.

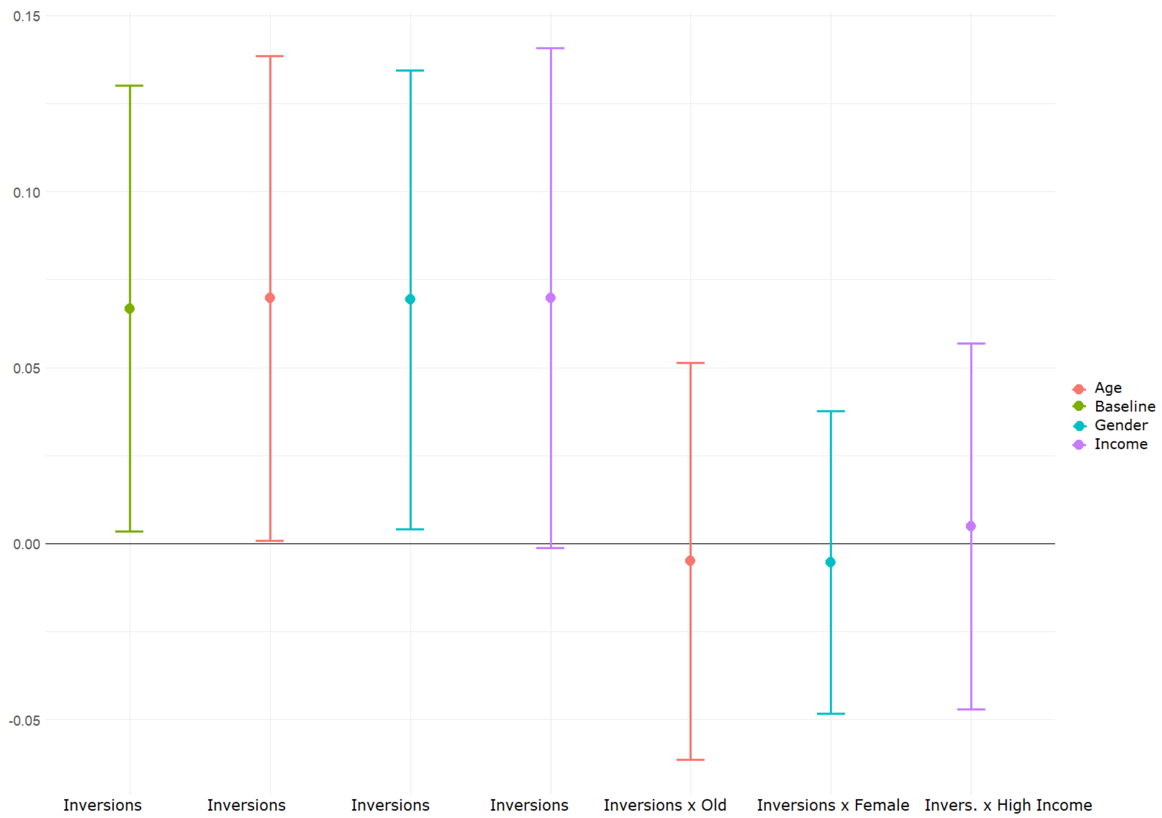


Figure 2.B.4.: Effect of thermal inversions on Neuroticism: Heterogeneity

Note: The figure depicts estimated coefficients from OLS regressions of Neuroticism on the share of days with a nighttime inversion during the in-utero period and interactions between inversions and child characteristics. Colors represent different regression models. Green: Baseline model (cf. table 4, Panel B, column 5). Red: Baseline model + interaction btw. inversions and dummy indicating age group 9-10. Blue: Baseline model + interaction btw. inversions and dummy indicating female child. Purple: Baseline model + indicator for above median income + interaction btw. inversions and high income indicator. All regressions include year-, month- and county-of-birth fixed effects, age group fixed effects, controls for parental education, child gender, migration background, age and age squared, cubic functions of temperature, wind speed and precipitation, plus one lead and one lag of the instrument. Bars represent 95%-confidence intervals, based on standard errors clustered at the county level.

3. Circadian Rhythms and Knowledge Worker Performance

Joint with Felix Holub and Ingo E. Isphording

3.1. Introduction

Despite the many technological and societal advancements of modern times, humans remain fundamentally governed by the circadian rhythm, the internal biological clock that regulates various physiological processes. A misalignment of the internal circadian clock and the social clock, often called *social jetlag*, causes sizeable economic and health costs (Giuntella and Mazzonna, 2019). Sleep deprivation, a widespread consequence, is estimated to generate costs of 1-3% of national GDP in OECD countries (Hafner et al., 2016). Increased levels of remote work and collaboration between workers across countries and time zones, driven by improved communication technologies and accelerated during the Covid19 pandemic, may further broaden the detachments of work time from natural and solar cycles, thus exacerbating these issues among affected workers. While an individual's internal circadian clock is largely predetermined, the social schedule is (to some degree) a policy choice: Policy-makers decide about abolishing or introducing daylight saving time, i.e., changing social clock time by one hour twice a year, or about later school start times, and firms decide about flexibility of work schedules. Therefore, the productivity consequences associated with social jetlag are important to quantify.

In this paper, we analyze how a misalignment of circadian clock and social clock affects the performance of high skilled knowledge workers. Using data from GitHub, the world's largest coding platform, we first provide descriptive evidence on performance differences between software developers based on their chronotype, i.e., their preferred sleep-wake cycles during the 24-hour day. Second, we use variation in timezone differences to collaborators over time, as well as DST transitions, to provide causal evidence on productivity effects of circadian misalignments.

We base our empirical analysis on data from *GitHub*, the world's largest code hosting platform, which provide time-varying proxies for performance of professional software developers. The data include precise timestamps of all activities conducted in public coding projects, allowing us to classify developers into morning- and evening-types. We focus on a sample of GitHub users who are highly active in company-backed coding projects since these users are likely professional software developers. We construct a developer-by-date panel including individuals on all continents and information on their total daily output quantity as well as activity levels during each hour of the day.

We classify developers into groups based on their chronotypes. Developers who are most active in the morning hours are classified as early chronotypes or *larks*, while developers who are most active in the evening hours are classified as late chronotypes or *owls*. We validate this classification by comparing variation in work quality and task complexity across hours of the day by chronotype, as well as further plausibility checks. Larks seem to strongly outperform later chronotypes by 23-29%. The direction of this substantial gap is in line with previous research on the productivity effects of chronotypes (Bonke, 2012; Conlin et al., 2022). Morning-type individuals are likely more productive than evening-type individuals because their natural rhythms are more aligned to prevailing social schedules (e.g., fixed work and school start times). We present evidence consistent with this in a new, more flexible high-skilled setting with global coverage. Yet, we cannot rule out that the chronotype classification picks up further characteristics that contribute to the performance gap, e.g., differences in the share of work in private GitHub repositories (and thus invisible to us).

While the evidence on performance differences by chronotype is descriptive and might be affected by selection, we next investigate whether misalignment of circadian clock and social schedules *causally* affect performance, using two complementary identification strategies.

First, we exploit within-user variation in time difference to collaborators which give rise to deviations between the internal body clock and work schedules. The data allow us to identify the collaborators of the developers in our panel, i.e., users working on the same coding projects during the same time period, as well as their time zones. We compute the average time difference to the collaborators for each developer and quarter. We show that developers work later in the day when they collaborate with users who are behind them on clock time, and earlier in the day when they collaborate with users who are ahead of them on clock time. This implies that there are incentives to synchronize work times within teams. We then show that the performance of larks drops when they start to collaborate with users who are behind on clock time (i.e., when they have the incentive to work late), while the performance of owls declines when they start to collaborate with users who are ahead on clock time (i.e., when they have the incentive to work early), relative to when the developers collaborate with users on similar clock time or a time shift in the more favorable direction for the respective chronotype. When working with collaborators whose local time is on average at least 2.5 hours shifted into the unfavorable direction, output declines by 5.7-6.3%.

Second, we exploit the natural experiment of transitions into and out of daylight saving time (DST), which creates exogeneous variation in developers' own social clock time independent of their collaborators. Moving into DST, i.e., setting clocks forward by one hour in spring, is equivalent to adopting the social clock time of the adjacent eastern timezone. It introduces a discrepancy between clock time and solar time which is a major determinant of the circadian clock. Using a regression discontinuity (RD) design, we find that the spring transition causes developers to start and end work slightly later, by 14 and 9 minutes, respectively, and causes output to drop by roughly 5.5%. By contrast, the fall transition, when clocks are set back by one hour and the original relation of social and solar time is restored, causes developers to start and end work earlier (13 and 10 minutes on average), and leads to an increase in output by 3.2%. This shows that even in a high-skill setting with a certain flexibility in work schedules, the DST policy generates declines in output and thus economic costs. When studying heterogeneity by chronotype, we find that larks are most affected.

Overall, our results imply that circadian rhythm disruptions have adverse impacts on the performance of knowledge workers. Individual workers and firms would likely benefit from measures to better accommodate workers with different chronotypes, especially owls. In line with previous research, we find that the widespread policy of DST generates relevant economic costs in terms of productivity losses.

Related Literature. A growing literature studies the causal impact of circadian rhythm disruptions and sleep duration on human capital. Exploiting the transitions into and out of DST, Smith (2016) and Jin and Ziebarth (2020), e.g., examine short-run impacts of sleep on fatal traffic accidents and health. Regarding long-run effects, Gibson and Shrader (2018), Giuntella et al. (2017), and Giuntella and Mazzonna (2019) exploit quasi-random differences in sunset time within time zones and at timezone borders that translate into later bedtimes, while wake up times are fixed due to social schedules. They find adverse impacts of shorter sleep duration on various health outcomes, cognitive performance and wages. Recently, using similar approaches based on sunset time, Costa-i Font et al. (2022) and Jagnani (2022) have shown that sleep affects wages in Germany and human capital accumulation among children in India. We add to this literature strand by replicating the analysis based on transitions into and out of DST in a different, highly relevant setting of high-skilled knowledge workers. Moreover, we study heterogeneity in the effects of the clock changes by chronotype to investigate which groups are most vulnerable to these disruptions in circadian rhythms.

A strongly related set of papers has studied the causes and consequences of the timing of activities. Hamermesh et al. (2008) for instance have found that the timing of sleep, TV watching, and working is affected by TV schedules, timezone (social clock), and, especially for workers in national industries, by a need to synchronize with others. Baylis et al. (2023) have recently shown that the timing of tweeting, visiting businesses, and departing to work are affected by local sunrise time, i.e., both the social clock and the solar clock are relevant in determining the timing of activities. Most strongly related to us, Chauvin et al. (2021) exploit data from a multinational enterprise and DST transitions to show that increases in temporal distance to co-workers induce workers who mainly conduct non-routine and managerial tasks to shift communication activities outside of standard business hours. Regarding the consequences of time of day, Gaggero and Tommasi (2022) have found that university students' performance in high-stakes exams depends on the time of test-taking, with a peak in performance around lunchtime, in line with the fact that young adults are typically not early chronotypes. By contrast, Pope (2016) shows that high school math test scores improve if math classes are scheduled earlier in the day. We contribute to this by providing direct evidence that coordination with co-workers in different time zones affects timing of work activity among software developers, i.e., synchronization is a relevant determinant of timing of activity in this group. Our main difference to Chauvin et al. (2021) is that we take into account how shifts in activity times affect circadian misalignment based on a developer's chronotype. We also show descriptively that performance of developers varies across hours of the day in line with their chronotype.

The comparison of output levels by chronotype also adds to a set of papers that investigate how morning- or evening preferences affect wages (Bonke, 2012; Conlin et al., 2022) as well as school performance (Goldin et al., 2020; Zerbini et al., 2017). In line with results from these studies, we document that early chronotypes outperform later types. In our large, international sample with high-frequency output measures, we have an advantage in external validity, and we can measure the performance gap separately on working days and free days. In line with the hypothesis that late chronotypes are disadvantaged because their natural rhythms are not aligned with social schedules, we find that early chronotypes outperform them only on working days.

3.2. Background

In this section, we provide background information on human chronotype, and why it might affect worker performance. We also describe current DST policies, which we exploit as an exogenous source of circadian disruptions.

Our internal biological clock synchronizes to the environmental 24-hour day, the light-dark cycle driven by the solar clock, in a process called entrainment. Individuals differ in how they align their circadian clocks to the external 24-hour day, i.e., in their phase of entrainment or chronotype. People with different chronotypes differ in preferred sleep times and in the time of day when they reach peaks in physical and cognitive capacity. In the population, chronotype approximately follows a normal distribution.[1] Early chronotypes are often called larks while late types are denoted as owls. Everyday life in most societies follows relatively rigid social schedules, e.g., fixed work and school start times, which are often well aligned with the natural rhythms of larks. Among late chronotypes, these schedules can give rise to *social jetlag*, a misalignment between internal and social clocks (Roenneberg et al., 2019a). Thus, worker performance and productivity might differ based on chronotype. Previous research has shown negative correlations between "lateness" and performance in school and university settings (e.g. Goldin et al., 2020; Zerbini et al., 2017; Zerbini and Merrow, 2017).

Chronotype is in part determined by genetic factors (e.g. Jones et al., 2016), but it also responds to natural cues. To synchronize to the external 24-hour day, the biological clock strongly relies on exposure to light and darkness, and the timing and strength of the light-dark cycle influences chronotype. In locations with an earlier solar clock (earlier average sunrise and sunset time), the chronotype distribution is earlier than in locations with later solar clocks (Roenneberg et al., 2007; Papatsimpa et al., 2021). Lastly, chronotype varies systematically across age and sex. During adolescence, people shift towards later chronotypes, with lateness peaking around age 20, and then continuously become earlier again. Before age 40-50, women are on average earlier types than men. Within periods of a few years and under fixed natural cues, chronotype is highly stable due to its genetic component (Fischer et al., 2017; Roenneberg et al., 2019a).

Objective biochemical measures like dim-light melatonin onset are the gold standard to assess chronotype, but expensive and impractical for large samples (Burgess et al., 2018). Frequently used alternatives are survey-based like the Morningness-Eveningness-Questionnaire (Horne and Östberg, 1976) that asks for example about preferred times of the day to carry out various activities, or the Munich ChronoType Questionnaire (Roenneberg et al., 2019a) that assesses chronotype by the mid sleep point on non-working days using self-reported sleep data. More recently, researchers have started to exploit data on online activities to study activity profiles and assess chronotype (Roenneberg, 2017; Smarr and Schirmer, 2018).

Daylight Saving Time (usually called Summer Time in Europe) refers to the practice of setting clocks forward by one hour in spring, and setting them back by one hour in autumn. This policy, which changes the *social* clock time, is observed in many countries, including the whole European Union, most of the US, large parts of Canada and Australia as well as other

[1] This refers to chronotype measured by the Munich ChronoType Questionnaire, a survey-based measure relying on self-reported sleep times (Roenneberg et al., 2019a).

countries like New Zealand and Chile. The clock changes occur during the night, e.g., at 1 am UTC in the European Union (i.e., at 1, 2, or 3 am local time) and at 2 am local time in the US. The precise transition days vary across countries and years, but mostly occur on weekends. In Europe, for instance, Summer Time lasts from the last Sunday in March to the last Sunday in October, whereas in the US, the spring transition falls on the second Sunday of March and the fall transition occurs on the first Sunday of November. In the Australian states that observe DST, it begins on the first Sunday in October and ends on the first Sunday in April.

Evidence from time use and survey data implies that the shortened night on the spring transition date causes US-Americans to sleep less (Barnes and Wagner, 2009), while the longer fall transition night leads to longer sleep duration (Jin and Ziebarth, 2020). Apart from changing the duration of the transition days, entering DST also leads to a reallocation of daylight from morning to evening, because social clocks are shifted forward relative to solar time. In fall, sunlight is reallocated back from evening to morning, as social clocks are moved back. Since circadian clocks entrain to sunlight this can also cause disruptions to circadian rhythms, e.g., if in spring people wake up after sunrise before the transition, and then wake up before sunrise again after the transition (Roenneberg et al., 2019b). In fact, Kantermann et al. (2007) show that sleep time on free days tracks sunrise during standard time, but this relationship is interrupted during daylight saving time.

Whether to maintain the biannual clock change or not, and in the case of stopping it, which time to implement permanently is a controversial issue. The parliament of the European Union voted to abolish DST in 2018, and the US Senate voted to permanently implement DST in 2022, but both votes still await further legislative action.

3.3. Setting and Data

GitHub. GitHub is a website that was founded in 2008 and allows users to store, manage and collaborate on coding projects. It is based on the version control system Git, a software that records which modifications were made to a file by whom at what point in time. GitHub is a web hosting service for Git projects, also called repositories (or repos for short). Repositories hosted on GitHub can be either public or private. Public repos, along with all associated files and the full history of activities, are visible to everyone, while private repos are only visible to the project owner and invited members.[2] In mid-2019, more than 120 million public repositories were hosted on GitHub by more than 30 million users, making it the world's largest host of source code. There are no limitations regarding who can host what types of projects

[2]Before 2019, only users with paid accounts were able to create private repositories, while free accounts were limited to using public repositories.

on GitHub. The most active projects in terms of contributions and subscriptions, however, are mostly owned or maintained by commercial tech companies.[3]

Users can conduct different types of activities on GitHub. The core action in Git is to take a snapshot of the current stage of the repository after implementing a change to a file, or a set of files, which is called a *commit*. Commits thus indicate that the commit author worked on code files and is saving their updated version. In addition to this version control functionality, GitHub provides features to facilitate collaboration between developers, most importantly pull requests and issues. A pull request (PR) is a tool to propose code changes to a repository. After a user has submitted a PR, it is reviewed by repository members who have to decide whether to accept (i.e., merge) or reject the potential changes. They can also provide feedback on the PR in an associated discussion forum. Issues are text messages, i.e., not directly related to coding. Common reasons to open an issue in a repository are reporting a bug, requesting a new feature, or organizing open tasks. As with PRs, it is possible to comment on issues to discuss the problem or question at hand. Once an issue is resolved, it can be marked as closed. While only repository owners and team members invited by them can modify files via commits, every GitHub user can open issues or send PRs to a repository. As all the described activities reflect work aimed at building software products, we collect data on them to measure output generated by software developers.[4]

Data. Our main data source is a database generated by the GHTorrent project, which provides information on all GitHub users, public repositories, and records of all actions conducted in these repos up to June 1st, 2019. Records of activities are available separately for different types of actions (commits, creating issues, PR comments, etc.) and include identifiers for the acting user and the repository the event was conducted in, as well as exact timestamps. In addition, the data contain information on all GitHub users, including their registration date as well as the location of residence and company affiliation, both of which users can optionally report on their profiles. The data also provide the names of all public repositories and the user id of the project owner (which can be individual or organizational users). Since the database only provides a snapshot of user information as of June 1st, 2019, and users might change locations over time, we complement this with user data from previous versions of the database providing one additional snapshot for each year from 2015 to 2018.

[3]Examples include Microsoft's VSCode (an open-source code editor), Meta's react (a JavaScript library for building user interfaces), or Google's TensorFlow (a machine learning framework). Potential advantages that companies can gain from engaging with the open source community are free labor provided by external contributors (e.g., reporting bugs in the software) and discovering talented developers who they might want to hire.

[4]GitHub also provides social network functions, e.g., options to follow other users and subscribe to specific repositories and issues to receive notifications about new activities, which we exploit to assess whether a user is still active on GitHub at all.

We combine this with data from GHArchive, which also provides records of all activities in public repositories. This data source contains additional information on PRs, namely the number of new lines of code added, the number of lines of code deleted, and the number of code files changed, which we use as proxies for PR complexity.[5]

An important caveat is that the data offers no information on activities in private repositories. Many GitHub users rarely or never work in public repos, such that the data we have neither allows to approximate their daily work output nor to determine phases of peak productivity during the day and thus the chronotype of these users. Keeping this limitation in mind, we next describe how we construct our analysis sample, intending to focus only on users who are professional software developers and do a substantial part of their formal work in public GitHub repos.

Sample Construction. To capture a sample of GitHub users who are professional software developers, we exploit the fact that many tech companies own public projects on GitHub. The core development teams of these projects likely comprise mainly company employees. We compile a list of company-backed repositories and identify users who are authorized to directly modify their files, i.e., users who made a commit in any of these repos at any point in time. From this sample, we drop users who do not provide information on their location of residence on their profile which allows us to determine the timezone in which they reside. In most cases, the country of residence is sufficient, but in a handful of large countries spanning multiple time zones, e.g., the US, Canada, or Brazil, more detailed information at the subnational level is required. The timezone information is needed to translate activity timestamps, which are reported in coordinated universal time (UTC), into local time to study temporal work patterns.[6] Additionally, we drop a small number of users who might be bots based on extremely high activity levels, very regular commit patterns, or suspicious login names.

Finally, to ensure that we capture only users who conduct a substantial part of their total work in public repositories, we impose an activity threshold they have to pass to be included in the sample: Users enter the sample from the month after they first made at least 20 commits in any given month. They remain in the sample until the end of the observation period unless they conduct less than three *unproductive* actions in a given month. We define unproductive actions as those that we do not include when measuring output. They are mostly based on the

[5]GHTorrent and GHArchive data can be linked based on users' login names and characteristics of the PRs (repository and number).

[6]We include users if we have suitable location information in any of the five snapshots covering the years 2015 to 2019. If no information is available during some of the years, we impute this with locations reported in earlier or later years. If users report different locations over time, we assume that they move as the new calendar year starts, i.e., in each calendar year, we include them with the location information from the respective year.

social network functionalities of GitHub like following other users or subscribing to repositories and issues. This way, we intend to remove users from the sample when they do not use GitHub at all, e.g., during vacations or when working on projects outside the public GitHub domain.

For the analysis at the individual user level, we get a final sample of 49,850 users across 154 countries, with the five most frequent countries being the US (18,200 users), the UK (3,200), China (3,100), Germany (3,000), and India (2,000). We observe all actions that these users conduct between January 2014 and May 2019, but, as explained above, the final user-by-date panel is unbalanced as we intend to include users only when they are sufficiently active in public repos. The user-by-date panel includes 31.1m observations, and on average, each user appears for 624 days.

To measure temporal work patterns, we exploit the precise timestamps reported in the data. We approximate the start, midpoint, and end of users' work day by the minutes of the day the first, average, and last action of the day were performed.[7] In addition, we record how many actions were conducted during each of the 24 hours of the day. We measure users' overall daily output by the total number of productive actions conducted, including commits, comments written on PRs, issues or commits, creations of PRs and issues, closing of PRs and issues, and reopening of issues. While output quantity is our primary outcome in most of the analysis, we also construct measures of output quality and task complexity which we will use to validate our chronotype classification. To assess the quality of users' output, we exploit the specific feature of pull requests: PRs are proposed code changes that can be accepted or rejected by the repository team. PR rejection points to issues in the code or style. Thus, a straightforward measure of output quality is the PR acceptance rate, defined as the share of all PRs opened on a given day that are merged.[8] For PRs, we are also able to assess task complexity by the lines of code added and deleted, as well as the files changed in a PR. For all three measures, we take the average across all PRs users worked on a given day, i.e., all PRs opened, closed, or commented on (reflecting coding and reviewing code). To generate a single complexity measure, we apply the inverse hyperbolic sine transformation to each of the three variables, standardize them, and take their average.

Descriptives. Table 3.1 shows summary statistics for our user-by-date sample. On average, users conduct 2.55 actions per day. The average user has a positive number of actions on 36%

[7] When creating these variables, we first shift all timestamps by four hours, such that we define a day as lasting from 4am to 4am of the following day.

[8] A PR can either get merged, get rejected (closed without merging), or can be left open. Since the project teams need some time to review a PR, we drop the final three months of the observation period when considering the PR acceptance rate. PRs that are not merged, but still open after less than three months might just not have been reviewed yet.

of all days. Conditional on any activity, the mean number of actions is 7.16. The first and last action of the day, on average occur at 13:21 and 17:49, respectively. Both measures show large variation, which suggests that different chronotypes might be present in the sample. Thus, our sample includes users who are highly active in public repositories, and for whom, conditional on being active at all, activity on the platform is likely a major share of their overall daily work. Hence, in this sample, the timing of GitHub activities throughout the day is likely informative about chronotype. In the following, we refer to the users in our sample as developers.

Table 3.1.: Summary Statistics

	$Q_{0.25}$	Mean	St. Dev	$Q_{0.75}$	Observations
Actions	0.0	2.55	6.7	2.00	31,123,125
$\mathbb{1}\{\text{Actions} > 0\}$	0.0	0.36	0.48	1.00	31,123,125
Actions | Actions > 0	2.00	7.16	9.68	9.00	11,091,412
Time of First Action	9:37	13:21	5.04 h	16:34	11,091,412
Time of Last Action	14:08	17:49	5.16 h	22:01	11,091,412
PR Acceptance Rate	0	0.64	0.40	1	2,320,731

Notes: Table depicts summary statistics for the developer-by-date panel based on the sample of 49,850 GitHub users we include in our analysis sample.

Collaboration Across Timezones. To analyse how coordination with collaborators in other timezones affects timing of work activities and performance, we measure for each developer in our sample how strongly they are connected to users in other timezones. We first identify their collaborators as users who work on the same projects during the same time period. We define two users as collaborators in a given quarter if both made at least 20 actions in shared repos. This way, we consider only important links that might plausibly induce coders to synchronize the timing of their activities. A limitation is that not all users report their location. We have to drop roughly 50% of those users that we identified as collaborators of our sample users because of missing location information. Next, for each sample developer i and quarter q, we compute the difference in local clock time to each collaborator j, $TimeDiff_{i,j,q}$, where positive values indicate that j is ahead on clock time (or located in a timezone east of i) and negative values indicate that j is behind on clock time (or located in a timezone west of i). Then we take a weighted average time difference across all connections $j \in C_{i,q}$, where $C_{i,q}$ denotes the set of users matching our definition of collaborators of i and providing valid location information. Weights are given by the number of activities in shared repositories $R(i, j, q)$ by the focal developer i:

$$TimeDiff_{i,q} = \frac{\sum_{j \in C_{i,q}} actions_{i,R(i,j,q),q} TimeDiff_{i,j,q}}{\sum_{j \in C_{i,q}} actions_{i,R(i,j,q),q}}$$

Due to missing location information and the fact that some developers conduct little work in shared repos, we have information on time difference to collaborators for a subsample of 36,330 sample developers and 214,440 developer-by-quarter observations.[9] In this sample we know the timezone of 65% of all collaborators on average and the mean timezone difference to the known co-workers is -9.2 minutes, i.e., sample developers are on average slightly ahead of their collaborators.

3.4. Chronotype and Performance

This section starts with a description of how we classify users by chronotype. This is followed by validation exercises to check that our classification is indeed driven by different circadian rhythms and not other user characteristics. Then, we measure the performance gap between different chronotypes.

3.4.1. Detecting Chronotype with k-Means Clustering

As mentioned above, chronotypes differ in their preferred sleep-wake rhythms and phases of peak physical and cognitive performance. In our sample of highly active GitHub users, times of the day of most intense GitHub use are a reasonable approximation to phases of peak productivity. Thus, we exploit the exact timestamps in the GitHub data, to measure during what parts of the day a user is most active to assess his chronotype.

We use a k-means clustering algorithm to classify users into different types, based on eight variables that measure what share of a user's total activity falls into each of eight three-hour windows of the day, $[0am, 3am)$, $[3am, 6am)$, $[6am, 9am), \ldots$. We choose $k = 3$ to separate users into early, intermediate and late chronotypes. Following previous work, we consider only activities on free days (weekends and public holidays) in the classification because work days are likely constrained by social schedules (e.g., work and school schedules) and thus less informative about chronotype (Bonke, 2012; Smarr and Schirmer, 2018). Thus, we restrict this analysis to users who are sufficiently active on free days (at least five free days with positive activity levels).

[9] In this subsample, we focus on developer-by-quarter observations where we know the timezones of at least 20% of all collaborators.

Figure 3.1 shows the output of the k-means clustering. The upper left plot displays activity patterns of the three types on work-free days. We detect three types with very distinct activity profiles, with peaks in the morning (10-11am), afternoon (3-4pm), and late evening (10-11pm), respectively. We denote these as larks, an intermediate type, and owls. The right plot depicts the activity patterns for the three chronotypes on work days. They are much more uniform and activity is strongly concentrated during standard working hours (8am-6pm) for all three types, pointing towards binding social schedules.[10] Nevertheless, we can see distinct patterns by chronotype, e.g., higher activities of larks during the early morning, and a second peak in activity in the evening/after a dinner break among owls. The bottom plot shows the prevalence of the different types in our sample. The distribution is relatively balanced, with the intermediate type appearing most frequent (38%) and larks being the least frequent type (27%). This is in line with findings in other samples (Roenneberg et al., 2019a).

In the appendix, we show that the differences in hourly activity shares across chronotype are statistically significant (Appendix Figure 3.A.1). We also repeat the k-means clustering based on activity patterns for the full sample of 49,850 users, using activity shares during the eight three-hour windows on all days, including working and work-free days (Appendix Figure 3.A.2).[11] In this extended sample, the share of different chronotypes as well as their activity patterns are similar to the sample with higher activity on free days.

3.4.2. Validation

Next, we validate that our classification is indeed driven by different circadian rhythms and not other user characteristics.

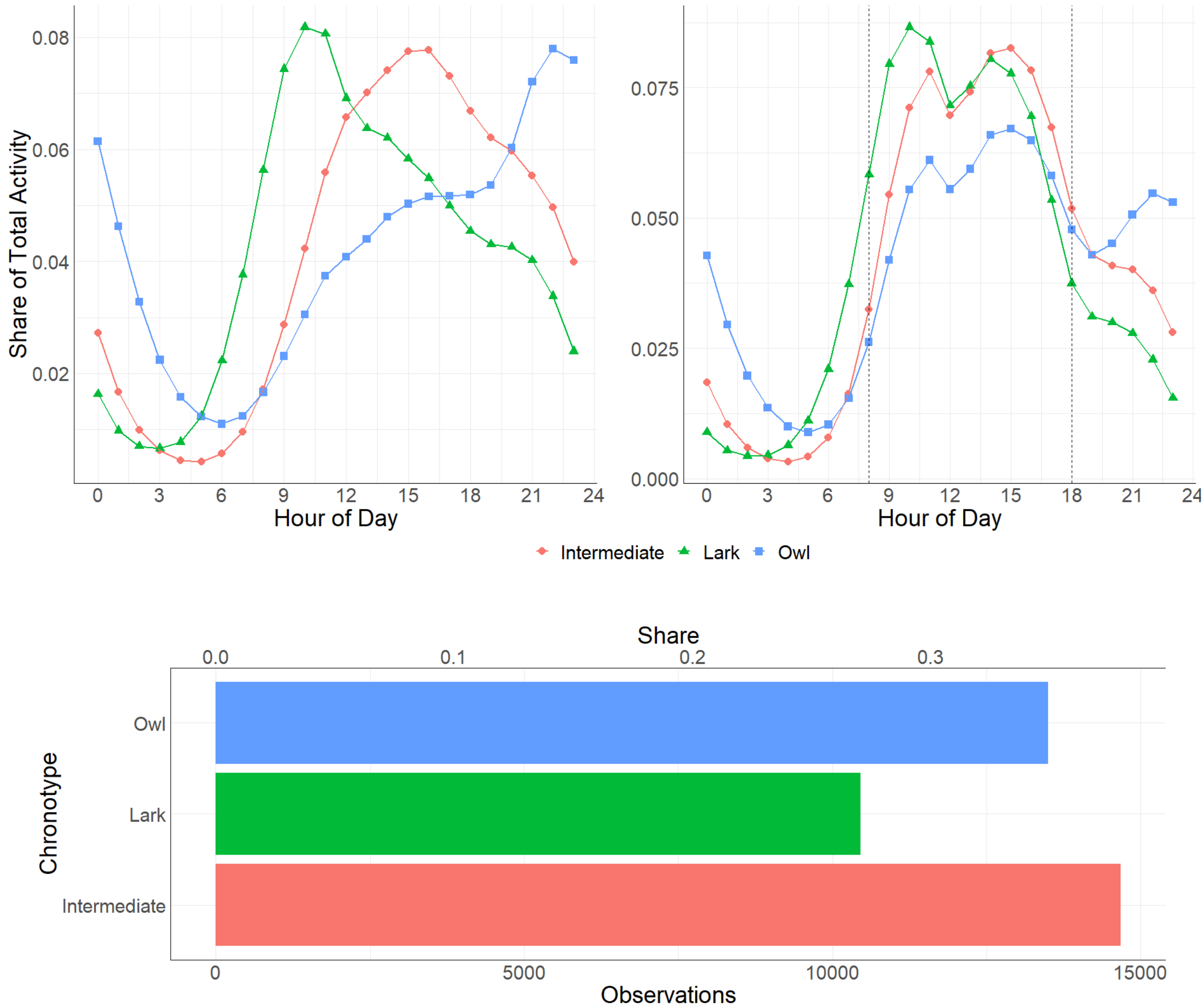
Alternative Explanations. We start by checking whether other developer characteristics are systematically correlated with chronotype and might be the true reason for the classification.

Early chronotypes might just be forced to get up early because of strict work schedules. Possibly, all developers would prefer to sleep late and work in the afternoon and evening, and some just appear like larks due to binding work schedules, whereas freelancers and self-employed are able to follow their natural rhythms. Also, users classified as owls might just work in public repos as a side or leisure project such that they mainly have time for this after standard working hours. These concerns are already greatly mitigated as we classify develop-

[10] We view this as supporting evidence that the GitHub users we include in our sample are indeed professional software developers.

[11] To take into account that temporal work patterns differ by type of day, we use 16 variables measuring the share of total activities during eight three-hour windows of the day, separately for work days and free days in the k-means clustering.

Figure 3.1.: Chronotypes Resulting From k-Means Clustering



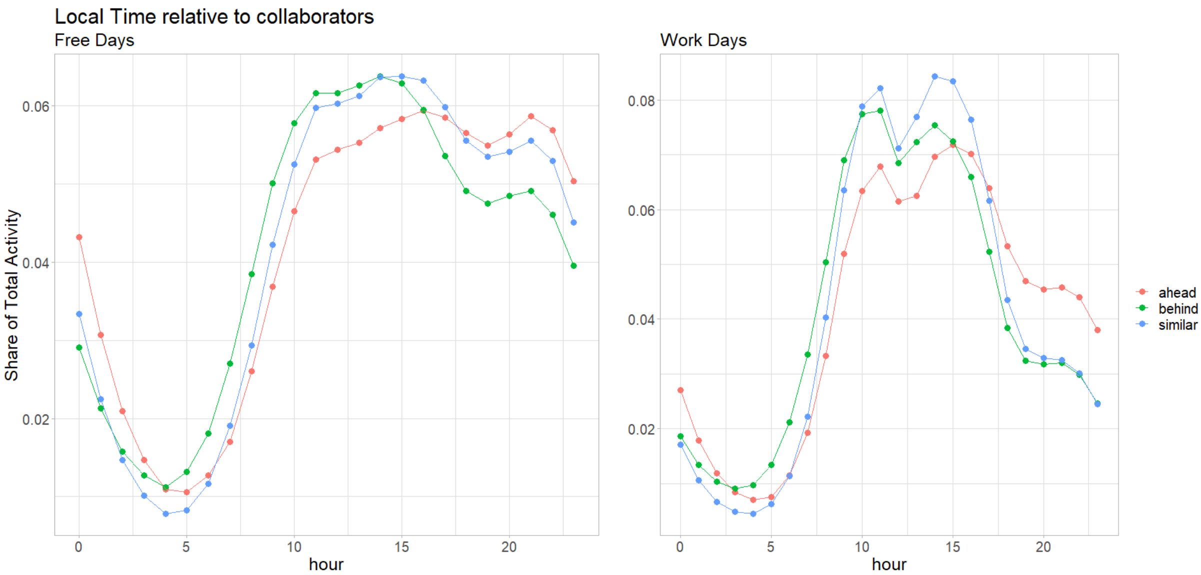
Notes: Top: Activity Patterns by three types resulting from k-means clustering algorithm, separately for weekends and public holidays (left) and for working days (right). Bottom: Number of users and share of total user sample by k-means type.

ers based on activities on free days only. However, if employees maintain work day patterns on the weekends due to habit formation, we might confound chronotype and employment status. To check this, in Appendix Figure 3.A.3 we depict employment status by chronotype. Larks are only slightly more likely to report to work at a company (as stated on user profiles in June 2019), with 65% compared to 60% among owls. The left plot shows the top employers. The share of larks working at these top tech companies is in most cases larger than the share of owls and intermediate types, but the later types are also present at the companies. Larks are also not substantially less likely to report to work as freelancers. This suggests that differences in chronotype as measured in the data do not simply reflect differences in employment status.

Alternatively, all developers might have a preference to work late, but those with children are forced to get up early because small children typically are early chronotypes or they have to be taken to childcare. Those larks might just reflect developers with children. The data includes no information on any demographics or household characteristics. To conduct an indirect test, we compare work patterns across standard working days and working days that fall into school holidays, separately by chronotype (Appendix Figure 3.A.4). If one chronotype would predominantly comprise parents, we would expect that school holidays affect this group more than the other groups.[12] This, however, seems not to be the case. During school holidays, we see a slight reduction in average activity levels among all three chronotypes. The size of the relative reduction is similar across groups, suggesting that parents are not over-represented in any of them.

Lastly, we might be worried that the differences in work timing arise due to a need for coordination in international teams spanning across multiple time zones rather than individual preferences. We use the collected information on time difference to collaborators to check this. We find that coordination in remote teams has a noticeable impact on the timing of activities (see Section 3.5 for details). Developers who are behind (ahead of) their collaborators on average work earlier (later) during the day. While the need for synchronization in teams

Figure 3.2.: Temporal Work Patterns by Time Difference to Collaborators



Notes: Left plot: Average activity patterns on weekends and public holidays, separately for three subsamples comprising developers and quarters during which developers are on average at least 2.5 hours *ahead* of their collaborators on local clock time, at least 2.5 hours *behind* them, or within 2.5 hours of their collaborators on local clock time. Right plot: Average activity patterns on working days for the same subsamples.

spanning multiple time zones affects activity timing, it is not driving the different activity patterns we detect by chronotype. First, most developers have co-workers who have the same

[12]The data on school holidays by country is obtained from Lai et al. (2022).

or very similar clock time: For 57% of user-by-quarter observations for which we could determine the timezone difference to collaborators, it is at most two hours. This share is very similar across all three chronotypes (60%, 55%, and 54% for the intermediate type, larks, and owls, respectively.).

Table 3.2.: Impact of Coordination with Co-Workers and Chronotype on Activity Timing

	$\text{MeanActivityTime}_{it}$		
	(1)	(2)	(3)
$\text{TimeDiff}_{i,q(t)}$	-2.52^{***}		-0.95^{***}
	(0.16)		(0.18)
$\text{TimeDiff}_{i,q(t)} \times \text{WorkDay}_{it}$			-2.01^{***}
			(0.19)
> 2.5 hours ahead of collaborators		20.52^{***}	
		(1.32)	
> 2.5 hours behind collaborators		-4.95^{***}	
		(1.31)	
Lark_i	-77.20^{***}	-77.58^{***}	-113.76^{***}
	(1.42)	(1.42)	(1.72)
Owl_i	63.45^{***}	63.04^{***}	76.96^{***}
	(1.69)	(1.69)	(1.66)
$\text{Lark}_i \times \text{WorkDay}_{it}$			45.07^{***}
			(1.55)
$\text{Owl}_i \times \text{WorkDay}_{it}$			-17.42^{***}
			(1.84)
Observations	6,534,330	6,534,330	6,534,330
Adjusted R^2	0.10	0.11	0.10

Note: Table displays estimated coefficients from Equation 3.1. Estimated at the developer-by-date level. Standard errors are clustered at the developer level and reported in parentheses. The regressor $\mathbb{1}\{TimeDiff_{i,q(t)} < -2.5\}$ implies that the developer is > 2.5 hours ahead of his collaborators, while $\mathbb{1}\{TimeDiff_{i,q(t)} > 2.5\}$ implies that he is > 2.5 hours behind his collaborators on clock time. *p<0.1; **p<0.05; ***p<0.01.

Second, Figure 3.2 shows that when we classify developer-by-quarter observations into three groups based on average time difference to co-workers (more than 2.5 hours ahead of co-workers, within 2.5 hours of co-workers, or more than 2.5 hours behind co-workers), we see some variation in their activity profiles, but it does not replicate the stark differences found between the three chronotypes. Third, Table 3.2 shows results from regressions of average activity time on the time difference to co-workers and on chronotype. We run a simple regression at the developer-by-date level:

$$\begin{aligned} MeanActivityTime_{it} = & \beta TimeDiff_{i,q(t)} + \gamma Lark_i + \delta Owl_i + \rho_{R(i)} WorkDay_{it} + \\ & \theta_{R(i),dow(t)} + \theta_{R(i),y(t),m(t)} + \theta_{country(i)} + \epsilon_{it} \end{aligned} \tag{3.1}$$

The average time of day of activities conducted by developer i on date t in minutes is regressed on the average time difference of developer i to his collaborators in the current quarter $q(t)$ in hours, as well as the chronotype of i (with the intermediate chronotype as reference group). We controls for differences in activity time between working days and free days, as well as region-by-day-of-week and region-by-year-by-month and country fixed effects. Even when comparing users with the same time offset to co-workers, we see a strong impact of chronotype on timing (Column (1)). When we replace the continuous measure $TimeDiff_{i,q(t)}$ with two indicators for developers being at least 2.5 hours behind or ahead of their collaborators on clock time, respectively, we find that the difference in mean activity time between these groups is only 25 minutes, whereas the difference between larks and owls exceeds two hours (Column (2)). Moreover, when interacting the chronotype and time difference regressors with a working day indicator, we find that chronotype has stronger impacts on activity timing on free days whereas the time difference to collaborators has stronger effects on working days (Column (3)), indicating that the first is about preferences while the second is about incentives to synchronize work activities.

Work Complexity and Quality Across Hours of the Day. Next, we show that not only output quantity varies across times of the day in a chronotype-dependent way, but also output quality and the complexity of tasks developers work on, which we did not use as an input in the k-means clustering algorithm.[13] We run a simple regression:

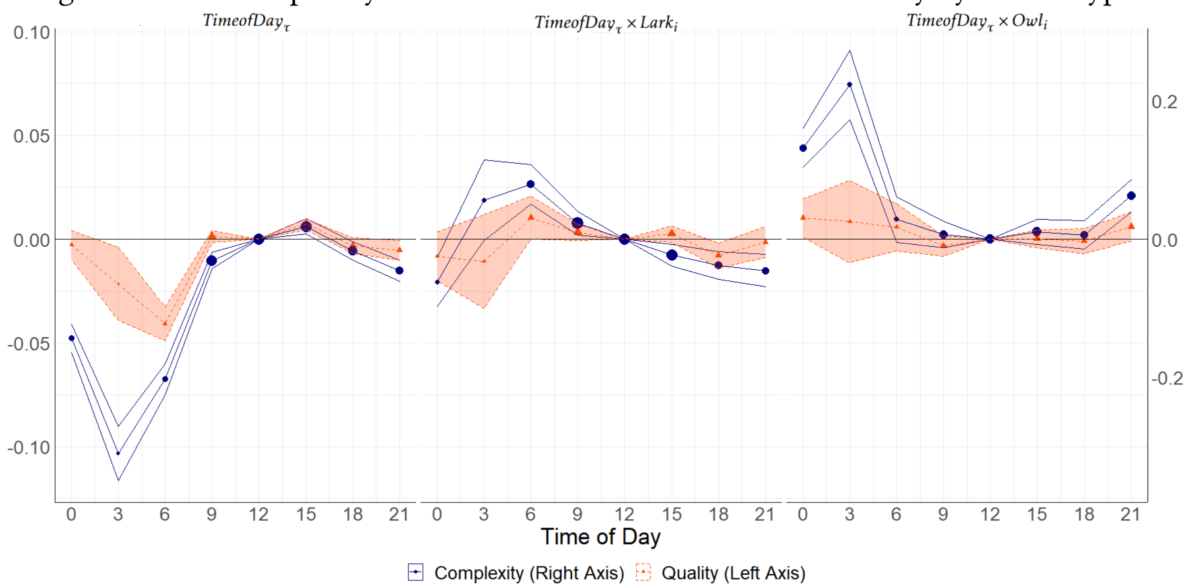
$$\begin{aligned} y_{i\tau} = & \alpha TimeofDay_\tau + \beta TimeofDay_\tau \times Lark_i + \gamma TimeofDay_\tau \times Owl_i + \\ & \theta_i + WorkDay_\tau + \epsilon_{i\tau} \end{aligned} \tag{3.2}$$

For the regression, the data is collapsed to the developer × time of day × type of day level, where time of day refers to three-hour windows, e.g., 0-3am (the same that were used to classify chronotypes), and type of day refers to working days vs. free days. $y_{i\tau}$ denotes outcome variables to capture complexity of tasks and quality of output. To measure output quality we

[13]Overall PR success rates vary between 63.6% and 65.1% for the three chronotypes.

use $PR\ Acceptance\ Rate_{i\tau}$, which measures what share of all PRs opened by developer i during time period τ get accepted. To assess task complexity, we use the index built based on the number of new lines of code added, lines of code deleted and distinct code files changed in PRs i worked on during time τ. $TimeofDay_\tau$ is a vector of seven dummies indicating the three-hour windows, with 12pm-3pm as omitted category. θ_i is a developer fixed effect and $WorkDay_\tau$ is a dummy indicating the type of day. The coefficients α reflect how average PR size or the likelihood of getting a PR merged varies across times of the day for the intermediate chronotype. The coefficients in vectors β and γ measure differences in these changes in outcomes across the day between the intermediate chronotype and larks and owls, respectively. The regressions are weighted by the number of PRs opened by developer i during time of the day τ and standard errors are clustered at the developer level.

Figure 3.3.: PR Complexity and Success Rate across Hours of the Day by Chronotype



Note: X-Axis: First hour of eight three-hour-intervalls of the day. Y-Axes: Magnitude of point estimates along with 95% confidence intervalls of α (left), β (center), and γ (right) in Equation 3.2. Blue circles and solid lines: Outcomes is an index of PR complexity, summarising number of new lines of code added, lines of code deleted, and code files changed (measured on the right y-axis). Red triangles and dashed lines: Outcome is the share of PRs opened that get merged (measured on the left y-axis). Point size reflects number of PRs opened during the respective hour by users of the respective chronotype.

Figure 3.3 shows that fluctuations in task complexity over the day for the different chronotypes are in line with expectations. The blue line indicates that developers with the intermediate chronotype work on larger PRs during the late afternoon relative to the baseline period, while they work on smaller PRs in the morning and evening (left plot). We also see large reductions in complexity during the night, however, these periods comprise relatively few activities. Relative to the intermediate type, larks work on larger tasks in the morning and smaller tasks in the evening (middle plot), whereas owls work on larger tasks in the late evening and night, between 9 pm - 2 am (right plot). The estimates depicted in red show similar patterns for the PR success rate. This confirms that the k-means clustering algorithm picks up biological differences in chronotype, because if, e.g., developers classified as larks were in fact forced to deviate from their natural rhythms and work early due to some external factors, we might rather expect that they pick easier tasks and perform worse than the other groups in the morning.

Response to Solar Time. Next, we show that our measures of chronotype depend on local solar time, in line with previous studies which found that chronotypes are on average earlier in locations with earlier average sunrise and sunset time (see Section 3.2).[14] We estimate a simple linear probability model at the level of developer i:

$$\mathbb{1}\{type = k\}_{il} = \beta sunrise_l + \gamma tenure_i + \theta_{L(l)} + \alpha pop_l + \epsilon_{il} \tag{3.3}$$

The outcome of interest is a dummy variable taking value one if developer i in location l has chronotype $k \in \{Lark, Intermediate, Owl\}$. The coefficent of interest, β, measures how the probability to be of type k differs between developers in locations with different average sunrise times (measured in hours). We control for developers' tenure on GitHub at the point when they enter the sample, $tenure_i$, the number of developers in location l, pop_l, and a geographic fixed effect, $\theta_{L(l)}$, defined as the interaction of country and latitude bin, or alternatively world region and latitude bin.[15] We control for pop_l as proxy of local population size because chronotype is on average later in urban as compared to rural areas (Roenneberg et al., 2019a). We exploit variation in average sunrise between locations in the same latitude range and within the same country or region, respectively. This variation arises because social time is fixed within time zones, but solar time varies based on longitude. Since we compute average sunrise time based on the days the user appears in the sample, additional variation can arise due to seasonal changes in sunrise time. Table 3.3 depicts the results. The number of observations in Panel A is reduced relative to the overall number of developers whose chronotype we could determine, because we only include developers with location information at the level of state or below to make sure that the measure of sunrise time is sufficiently accurate.[16] In Panel B we additionally drop developers who changed their location during the observation period. As expected, a later sunset time reduces the prevalence of larks and increases the prevalence of owls.

This serves as a further plausibility check, because it is unlikely that e.g., parents are system-

[14]We obtain data on sunrise time by longitude, latitude and date via the R package suncalc.

[15]World regions mostly follow the UN classification and comprise several countries, e.g. North America, South America, Western Europe, Eastern Europe, Southern Asia, Eastern Asia, etc.

[16]Sunrise time is measured at the centroid of the city or state the developer reports as location.

Table 3.3.: Effect of Local Sunrise Time on Chronotype

	Lark		Intermediate		Owl	
Panel A: *full sample*						
sunrise (hours)	-0.018^{**}	-0.022^{**}	0.001	-0.001	0.017^{*}	0.023^{***}
	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.007)
Observations	33,049	33,049	33,049	33,049	33,049	33,049
Panel B: *without movers*						
	-0.015	-0.020^{**}	-0.002	-0.006	0.017	0.026^{***}
	(0.009)	(0.010)	(0.009)	(0.010)	(0.010)	(0.008)
Observations	27,222	27,222	27,222	27,222	27,222	27,222
$\theta_{L(l)}$	Region	Country	Region	Country	Region	Country

Notes: The table depicts coefficient estimates of β in Equation 3.3. Standard errors clustered at the region x latitude bin level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01.

atically more likely to live in locations with later sunrise times than users without children.

3.4.3. Performance Differences by Chronotype

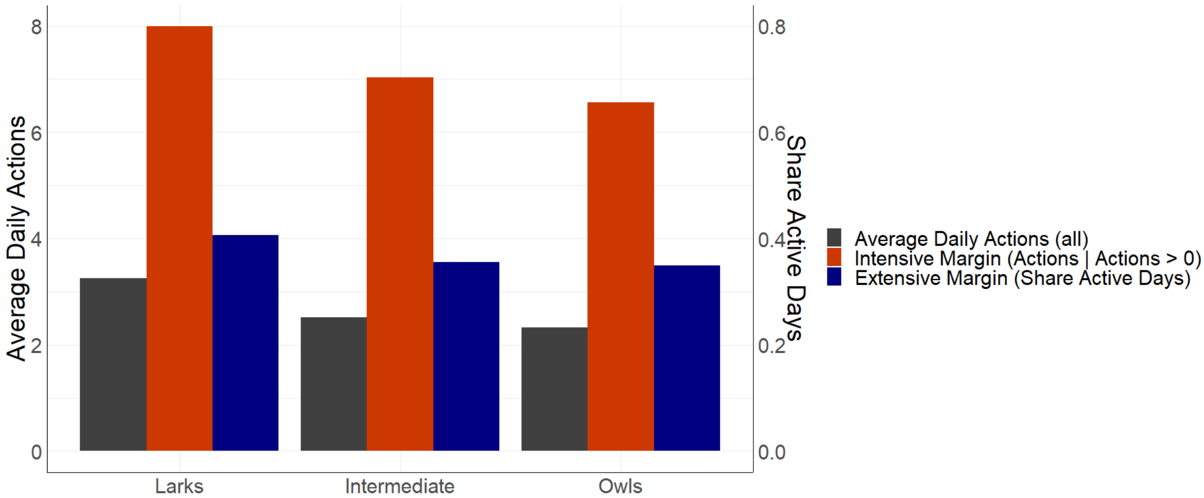
This section presents results on the overall gap in output quantity across chronotypes. Figure 3.4 shows the average number of daily actions by type in grey. Larks on average conduct 3.25 actions per day, whereas the intermediate chronotype and owls, respectively, only complete 2.51 and 2.32 actions, respectively, or 77% and 71% of the level achieved by larks. The red and blue bars assess the contributions of the intensive and extensive margin to this overall gap, showing the average number of actions on active days (red), and the share of total days during the sample period with positive activity levels (blue). We find that larks achieve higher levels on both margins.

In addition to this raw gap in daily output, we also estimate conditional performance gaps, using a simple regression of daily actions on chronotype indicators and several controls:

$$\begin{aligned} Actions_{it} =& \beta^L Lark_i + \beta^O \mathrm{Owl}_i + \alpha_{R(i)} h_{it} + \theta_{R(i),dow(t)} + \\ & \theta_{R(i),y(t),m(t)} + \theta_{country(i)} + \gamma\, tenure_{it} + \delta\, tenure_{it}^2 + \epsilon_{it} \end{aligned} \tag{3.4}$$

The estimates $\widehat{\beta^O}$ and $\widehat{\beta^L}$ reflect the performance gap between the intermediate chronotype and larks or owls, respectively, conditional on experience levels captured by $tenure_{it}$ and its square, region-specific variations in productivity across days of the week, $\theta_{R(i),dow(t)}$, region-specific time shocks, $\theta_{R(i),y(t),m(t)}$, and an indicator for public holidays, h_{it}. Moreover, by controlling

Figure 3.4.: Mean Activity Levels by Chronotype



Notes: X-axis: Chronotype. Grey: Average Daily Actions. Red: Average Daily Actions conditional on positive activity (both measured on the left y-axis). Blue: Share of days with positive number of actions during the sample (measured on the right y-axis).

for country of residence fixed effects $\theta_{country(i)}$, we take into account that both productivity and chronotype composition might vary between countries.

Results in Table 3.4 show that also conditional on these controls, larks clearly outperform the other types (Column (1)). The gap in total actions implies an advantage of 26.3% relative to average actions among the intermediate type and 30.5% relative to owls. In Appendix Table 3.B.1 we repeat this regression on several subsamples. When restricting the analysis to days with positive activity levels (intensive margin), developers with above median activity levels, developers reporting a company on their profile in 2019, or only developers residing in the US, $\widehat{\beta^O}$ and $\widehat{\beta^L}$ are of very similar magnitude.

The performance gap between larks and later chronotypes is in line with findings from previous studies (e.g. Bonke, 2012; Conlin et al., 2022; Zerbini et al., 2017).[17] As mentioned in Section 3.2, late chronotypes are likely to suffer from *social jetlag* because public schedules force them to deviate from their natural rhythms. To assess whether this is the reason underlying the performance gap, we exploit our multinational panel data to measure differences in activity in subsamples with more or less binding social schedules. In Column (2) of Table 3.4, we explore performance differences between chronotypes separately for working days vs. work-free days. The column depicts estimates from the following regression:

[17] Bonke (2012) and Conlin et al. (2022) have found gaps of 5-8% in wages in the general population conditional on several demographic controls.

Table 3.4.: Chronotype and Daily Output Quantity

	Actions$_{it}$	
Lark$_i$	0.661***	0.916***
	(0.056)	(0.069)
Owl$_i$	−0.048	−0.102*
	(0.045)	(0.053)
Lark$_i$ × Free day$_{it}$		−0.823***
		(0.054)
Owl$_i$ × Free day$_{it}$		0.177***
		(0.037)
Free day$_{it}$		−1.016***
		(0.027)
Observations	29,436,857	29,436,857

Notes: Table displays estimated coefficients from Equation 3.4 in the left column, and from Equation 3.5 in the right column. Standard errors clustered at the developer level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01.

$$\begin{aligned} Actions_{it} =& \beta^L Lark_i + \beta^O Owl_i + \eta^L Lark_i \times FreeDay_{it} + \eta^O Owl_i \times FreeDay_{it} + \\ & FreeDay_{it} + \theta_{R(i),dow(t)} + \theta_{R(i),y(t),m(t)} + \theta_{country(i)} + \gamma tenure_{it} + \delta tenure^2_{it} + \epsilon_{it} \end{aligned} \tag{3.5}$$

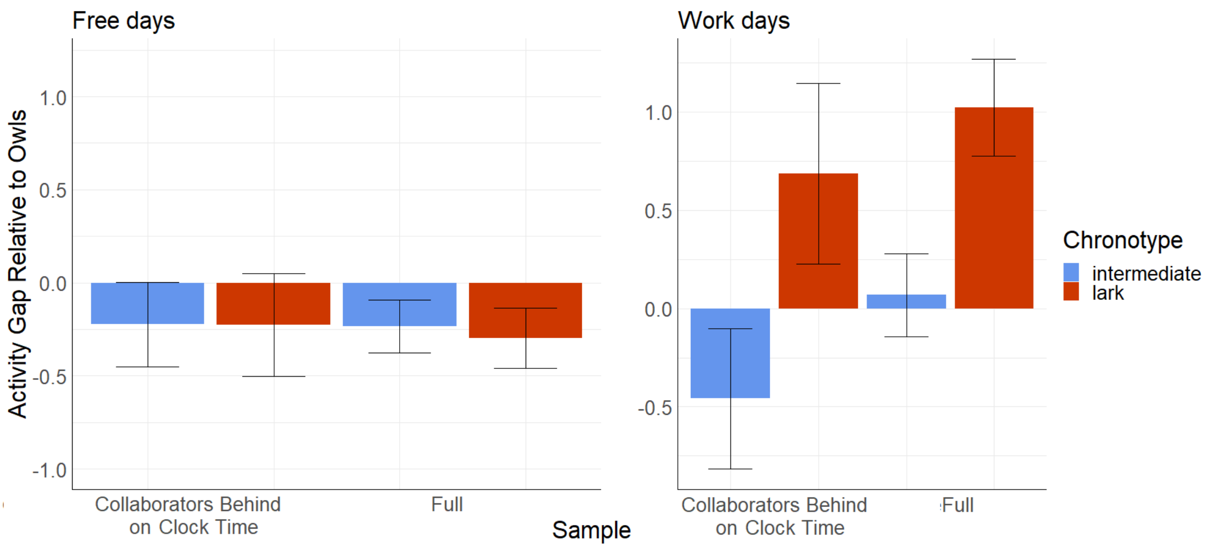
Free days, i.e., weekends and holidays, are less constrained by social schedules than standard working days. The gap in activity levels between larks and the other types amounts to .92 and 1.02 on work days for the intermediate type and owls, respectively, which is even larger than in the full sample, but it is small on free days.[18] This suggests that the gap arises because social schedules are well aligned with natural rhythms of larks, but disrupt the circadian rhythms of later chronotypes.

Next, we exploit the fact that we have information on collaborators and their time zones. We compare performance between chronotypes among developers with different timing incentives due to the composition of their collaborators. As shown in detail in the next section (Table 3.5 and Figure 3.6), developers with collaborators who are, on average, ahead of them on clock time tend to work earlier while developers with collaborators who are behind them

[18]To put this into perspective, the average number of actions on work days are 2.64, 2.93 and 3.95 for owls, the intermediate type and larks respectively, and 1.60, 1.57 and 1.70 on free days.

on clock time tend to work later. Hence, owls with collaborators who are behind them on clock time might be able to work more in line with their natural rhythms because their natural activity time (late in their day) overlaps with standard work times in more western time zones. Consequently, these owls are likely less prone to social jetlag.

Figure 3.5.: Activity Gap between Owls and other Types by Time Difference to Collaborators



Notes: Plot depicts conditional gaps in daily actions between owls and the other chronotypes, along with 95% confidence intervals, separately for free days and work days. Performance gaps are estimated by estimating Model 3.5 on the sample of all developer-by-quarter observations with information on time zones of collaborators or only the subsample where collaborators are on average at least 2.5 hours behind on local clock time. Sample is indicated on the x-axis.

Figure 3.5 shows the performance gap between larks and owls (red) as well as between the intermediate type and owls (blue). The gaps are depicted separately for the full sample of developers with information on collaborators, and in the subsample of those whose collaborators are on average in more western time zones (behind them on clock time), or more formally the subsample with $TimeDiff_{i,q(t)} < -2.5$ (see x-axis).[19] Developers in this subsample have an incentive to work later to coordinate with their collaborators, i.e., more in line with natural rhythms of owls. The performance gap between owls and the other two types on *free days* is of very similar magnitude in the two samples. Owls conduct slightly more actions than developers of the other chronotypes in both samples (left plot). On *working days*, by contrast, the performance of owls relative to both other types is indeed better in the subsample of developers collaborating with users who are behind them on clock time than in the full sample (right plot). In the subsample, the advantage of larks relative to owls is smaller (red bars) and owls even outperform the intermediate type (blue bars). This suggests that misalignment of standard work schedules and the circadian clock gives rise to the overall performance gap.

[19]These activity gaps are obtained by estimating Model 3.5 on the respective samples.

3.5. Causal Impacts of Circadian Misalignment

The evidence presented so far suggests that deviations from natural rhythms due to social schedules have detrimental effects on the performance of late chronotypes. However, the underlying analysis is mostly descriptive as we exploit cross-sectional variation in chronotype and we cannot fully rule out that it might pick up other unobservable differences, e.g., whether work in public GitHub repos constitutes the main job or a side project. If social jetlag is indeed the underlying mechanism for the performance gap, we would expect that other factors that cause a mismatch between work time and the circadian clock also lead to a decline in output, irrespective of chronotype. In this section, we investigate this using two complementary identification strategies.

3.5.1. Impacts of Collaboration Across Time Zones

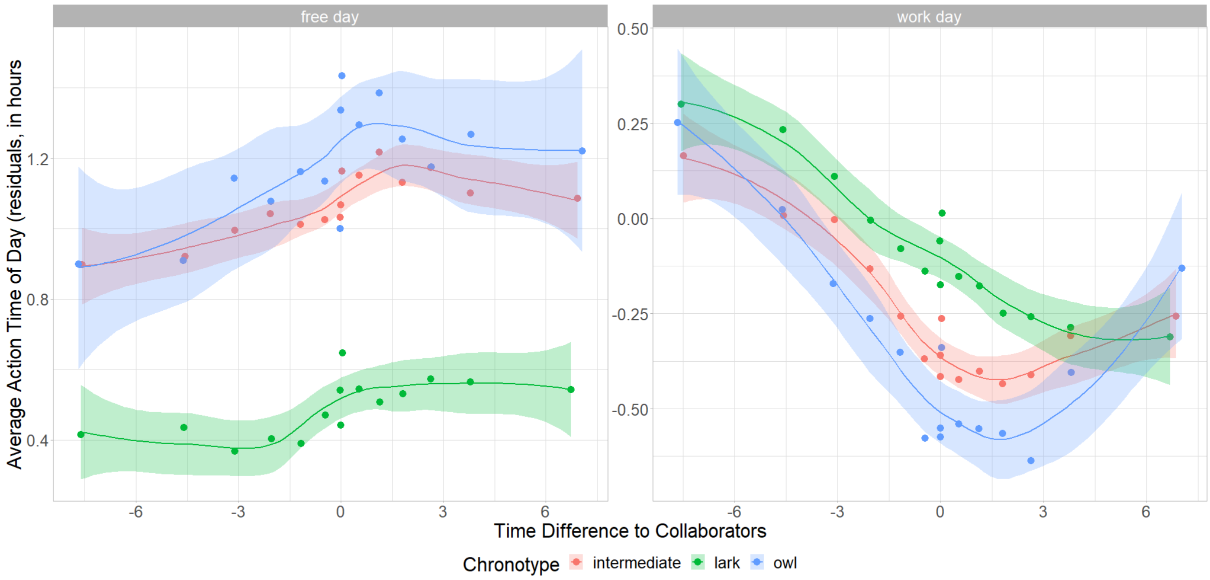
In our first approach, we exploit the fact that synchronization with collaborators in other time zones provides incentives for developers to adjust their work times. Figure 3.6 depicts the relationship between the time of day of the average action (in hours) and the average difference in clock time between the developer and his collaborators, separately by chronotype and for working and free days. Because both work schedules and average time differences to developers might differ across regions, we show binned scatterplots of residualized average daily activity time after demeaning by region and chronotype.[20] On work days, residualized average activity time is later if developers collaborate with users who are on average behind them on local clock time (negative time difference) and earlier if collaborators are on average ahead of them on local clock time (positive time difference). This correlation holds for all three chronotypes. On free days, by contrast, there is no systematic pattern in the relationship between time difference and work time. This suggests that the relationship on work days is driven by a need to synchronize work activities: If collaborators are ahead on clock time, a developer can facilitate coordination with them by working earlier than usual in his timezone to ensure that his work time overlaps with standard working hours at the collaborators' timezone.

While Figure 3.6 uses both within- and across-developer differences in collaborator composition, Table 3.5 exploits only within-developer variation in time differences to co-workers over time. It is based on a simple regression:

$$y_{it} = \beta \text{TimeDiff}_{i,q(t)} + \theta_i + \theta_{R(i),y(t),m(t)} + \theta_{R(i),dow(t)} + \epsilon_{it} \tag{3.6}$$

[20] Time difference to collaborators is very likely to differ across world regions. Developers in North America, for instance, are likely to collaborate mostly with users who are ahead on clock time (e.g., in Europe) due to their western location, while developers in Asia are more likely to collaborate with users who are behind them on clock time.

Figure 3.6.: Time Difference to Collaborators and Average Activity Time



Note: Binned scatter plot, based on developer × date observations with positive activity levels. Y-axis measures mean daily activity time, after demeaning by region and chronotype. X-axis measures average difference in local clock time between the developer and their collaborators in the current quarter. Positive (negative) timezone difference implies that collaborators are ahead (behind) on clock time. The relationship is depicted separately for three chronotypes and for working and free days.

where y_{it} represents three different outcomes measuring work timing of developer i on date t, $TimeofFirstAction_{it}$, $TimeofMeanAction_{it}$, and $TimeofLastAction_{it}$, each measured in minutes. For a one-hour increase in time difference, first, average, and final activity of the day become earlier by roughly 1 minute on free day and 2 minutes on work days. When they are more than 2.5 hours ahead of their collaborators on average, developers' work time is on average 10 minutes later than during periods when collaborators are on similar clock time. Conversely, when they are more than 2.5 hours behind their collaborators, developers work on average 8 minutes earlier. While these effects are moderate in size, we have to account for the measurement error in the regressors due to missing location information which likely attenuates the estimates.

Given these results, we can use *within-developer* variation in the composition of collaborators over time to study how deviations in work time from natural cycles affect performance. We expect that there will be an adverse impact on performance if larks collaborate with users who are behind on clock time such that they have an incentive to work late to synchronize, i.e., to deviate from the natural circadian rhythm. We expect that owls' performance, by contrast, is adversely affected when they start to collaborate with users who are ahead on clock time such that there is an incentive to work early to coordinate. To test this, we use the following model for owls:

Table 3.5.: Impact of Time Difference to Collaborators on Temporal Work Patterns

	Time of First Action$_{it}$			Time of Mean Action$_{it}$			Time of Last Action$_{it}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TimeDiff$_{i,q(t)}$		−1.82***	−1.12***		−1.79***	−0.93***		−1.49***	−0.71***
		(0.15)	(0.23)		(0.13)	(0.19)		(0.15)	(0.22)
TimeDiff$_{i,q(t)}\times$ WorkDay$_{it}$			−0.90***			−1.10***			−1.00***
			(0.24)			(0.20)			(0.21)
$\mathbb{1}\{$TimeDiff$_{i,q(t)} < -2.5\}$	9.46***			11.01***			10.75***		
	(1.10)			(0.90)			(1.11)		
$\mathbb{1}\{$TimeDiff$_{i,q(t)} > 2.5\}$	−8.34***			−7.95***			−6.52***		
	(1.07)			(0.91)			(1.17)		
Observations					4,784,611				
Unique Users					13,530				
Adjusted R^2	0.16	0.16	0.16	0.17	0.17	0.17	0.16	0.16	0.16

Note: Results from regressions of minute of day of first action (Columns (1) to (3)), average action (Columns (4) to (6)) or last action (Columns (7) to (9)) of the day on measures of average difference in clock time between the developer and his collaborators (during relevant quarter). Regressors of interest are either the time difference measured in hours, TimeDiff$_{i,q(t)}$, (Columns (2), (5), (8)), the time difference in hours and its interaction with an indicator for working day (Columns (3), (6), (9)), or two indicators for TimeDiff$_{i,q(t)} < -2.5$ hours or TimeDiff$_{i,q(t)} > 2.5$ hours (Columns (1), (4), (7)). Negative time difference implies that collaborators are behind on clock time, positive time difference implies that collaborators are ahead on clock time. All regressions control for user fixed effect, region-by-day-of-week fixed effect, region-by-year-by-month fixed effect and an indicator for work-free days. Estimations are based on subsample of 13,530 developers with variation in indicators used in Column (1) during the observation period. Standard errors clustered at the developer level. *p<0.1; **p<0.05; ***p<0.01.

$$\begin{aligned} Actions_{it} = {} & \beta_{O1}\mathbb{1}\{TimeDiff_{i,q(t)} > 2.5\} + \beta_{O2}\mathbb{1}\{TimeDiff_{i,q(t)} > 2.5\} \times WorkDay_{it} + \\ & \alpha WorkDay_{it} + \gamma tenure_{it} + \delta tenure_{it}^2 + \theta_i + \theta_{R,dow(t)} + \theta_{R,y(t),m(t)} + \epsilon_{it} \end{aligned} \tag{3.7}$$

And a similar model for larks:

$$\begin{aligned} Actions_{it} = {} & \beta_{L1}\mathbb{1}\{TimeDiff_{i,q(t)} < -2.5\} + \beta_{L2}\mathbb{1}\{TimeDiff_{i,q(t)} < -2.5\} \times WorkDay_{it} + \\ & \alpha WorkDay_{it} + \gamma tenure_{it} + \delta tenure_{it}^2 + \theta_i + \theta_{R,dow(t)} + \theta_{R,y(t),m(t)} + \epsilon_{it} \end{aligned} \tag{3.8}$$

The dummy variable $\mathbb{1}\{TimeDiff_{i,q(t)} > 2.5\}$ takes value one when the collaborators of i during the current quarter $q(t)$ are, on average, more than 2.5 hours ahead of i on clock time, whereas $\mathbb{1}\{TimeDiff_{i,q(t)} < -2.5\}$ indicates whether they are more than 2.5 hours behind on clock time. Since collaborator composition can change over time, so do these regressors for a given developer. Other covariates are as described above.

Table 3.6 depicts the results. When owls collaborate with users who are ahead of them in terms of clock time, they produce less output per day than when they collaborate with users on similar or later clock time. Larks, by contrast, perform worse when they collaborate with users who are behind on clock time as compared to when they work with users on similar or earlier clock time. The reductions only occur on work days, consistent with the fact that coordination with co-workers is less important on free days. Relative to the average activity levels on work days, the work day reductions amount to 6.3% for owls and 5.7% for larks. These results further suggest that deviations from preferred circadian cycles reduce knowledge worker performance.

Table 3.6.: Impact of Time Difference to Collaborators on Daily Activity by Chronotype

	$Actions_{it}$	
$\mathbb{1}\{\text{TimeDiff}_{i,q(t)} < -2.5\}$	−0.061	
	(0.113)	
$\mathbb{1}\{\text{TimeDiff}_{i,q(t)} < -2.5\} \times \text{WorkDay}_{it}$	−0.326**	
	(0.161)	
$\mathbb{1}\{\text{TimeDiff}_{i,q(t)} > 2.5\}$		0.190*
		(0.106)
$\mathbb{1}\{\text{TimeDiff}_{i,q(t)} > 2.5\} \times \text{WorkDay}_{it}$		−0.547***
		(0.108)
WorkDay_{it}	3.050***	2.098***
	(0.080)	(0.077)
Sample	Larks	Owls
Observations	2,918,963	2,433,618
Mean Dep. Variable on work days	6.78	5.65

Notes: Table depicts OLS estimates of coefficients from Models 3.7 and 3.8. Standard errors clustered at the developer level are reported in parentheses. *p<0.1; **p<0.05; ***p<0.01.

3.5.2. Impacts of Daylight Saving Time Transitions

In our second approach to study whether there is a causal effect of circadian disruptions on worker performance, we exploit a natural experiment, the transition into and out of DST, following e.g. Smith (2016). Previous research indicates that especially the spring transition causes a disruption to the circadian rhythm due to the one-hour loss during the night, which translates at least partly into shorter sleep duration, and due to the deviation of social time from solar time.[21] Evidence on the fall transition is mixed, Smith (2016) finds no effect on traffic accidents, while Jin and Ziebarth (2020) find positive impacts on health outcomes and reductions in daytime sleepiness. We investigate how the transitions affect performance of knowledge workers who typically have quite a lot of flexibility in their jobs, for example regarding work schedules. We not only consider a very different setting as compared to previous research, but we also have data on both daily output and proxies for daily work start and end time, namely the minute of the day when the first action and last action were conducted. We are thus able to investigate how strongly workers adapt to the clock change by changing working hours (as shown on the clock in local time).

For the analysis, we reduce the sample to only those countries and states that observe DST. Before estimating the transition effects, it is important to consider that the spring transition day has only 23 hours and the fall transition day has 25 hours, so there could be mechanical impacts on daily output just because developers have more/less time to work. Since we define a day as lasting from 4am to 4am, we circumvent this problem because the clock changes occur before 4 am in local time. This means that we assign the missing hour to the Saturday before the spring transition and the additional hour to the Saturday before the fall transition. This works against finding the hypothesized negative effect of the spring transition and the possible positive effect of the fall transition. Thus, we make no further adjustments for the differences in day length.

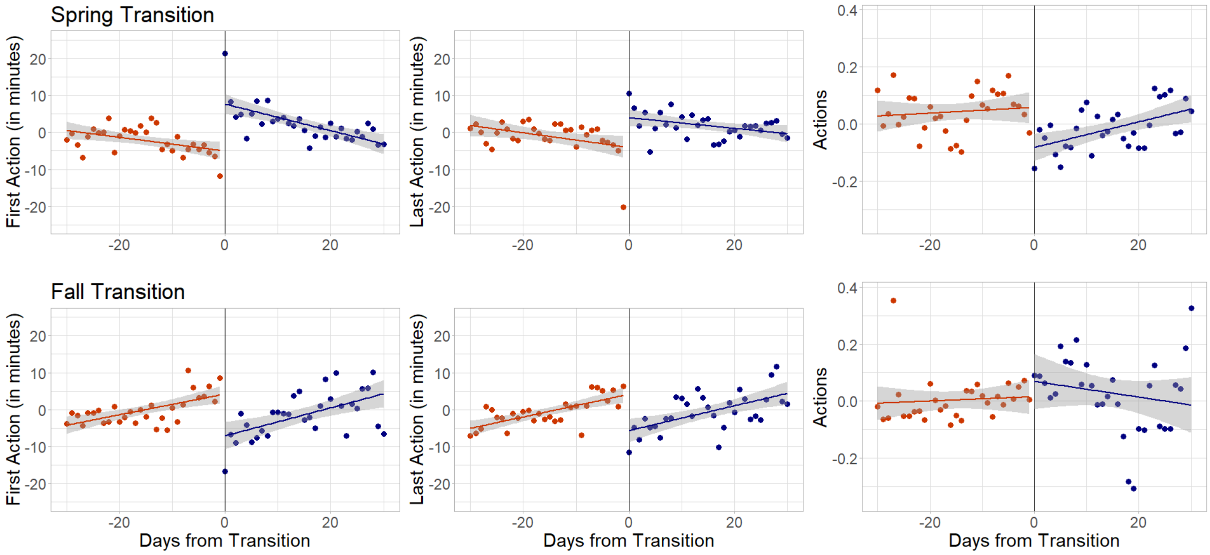
We residualize the outcomes of interest, number of total actions, minute of day of the first action and minute of day of the last action, with respect to developer-, region-by-day of week-, region-by-year- and region-by-month fixed effects as well as an indicator for public holidays. For a first visual inspection, we plot the average residuals against the time relative to the transition date (Figure 3.7). Regarding work times, we observe that both the time of the first and last action tend to become earlier before and after the spring transition, indicating that work time tracks sunrise.[22] Right at the transition, there are clear upward jumps in the time of first and last action, indicating that developers adjust to the clock change. However, the adjustment is only partial, as the discontinuities are clearly smaller than 60 minutes. At and around the fall transition, we see the opposite picture: Before and after the clock change, work times gradually become later in line with delaying sunrise times in fall, while at the transition there are clear negative jumps, i.e. work start and end are shifted earlier, but by less than an hour. These results indicate that both the social clock and the solar clock are relevant in determining working hours among knowledge workers.[23] Looking at residualized

[21]Switching to DST is equivalent to adopting the clock time of the adjacent eastern timezone, but without the corresponding change in solar time. Time zones are typically designed such that the clock time tracks solar time. Under DST, clock time is ahead of solar time.

[22]We present direct evidence on this relationship in Appendix Figure 3.A.5.

[23]This is in line with recent findings by Baylis et al. (2023) who show that a one-hour shift in solar time un-

Figure 3.7.: Residualized Daily Outcomes Plotted Against Time to DST Transition



Notes: Residuals are obtained by regressing the respective outcome on developer-, region-by-day of week-, region-by-year- and region-by-month fixed effects as well as an indicator for public holidays. Points reflect averages of all residuals for a given date relative to the transition date. Fitted lines with 95% confidence intervals are based on local linear regressions. Top: Spring Transition, bottom: Fall transition.

total actions, we find a downward jump at the spring transition, hinting at a decline in output. For the fall transition, the picture is less clear.

We use the following model to estimate the size of the discontinuities:

$$\begin{aligned} y_{it} =& \beta_0 + \beta_1 \mathbb{1}\{Days\ Since\ Transition \geq 0\}_{it} + f(Days\ Since\ Transition_{it}) + \\ & f(Days\ Since\ Transition_{it}) \times \mathbb{1}\{Days\ Since\ Transition \geq 0\}_{it} + u_{it} \end{aligned} \tag{3.9}$$

where y_{it} reflects the residualized outcomes mentioned above, measured at the level of developer i by date t. $DaysSinceTransition_{it}$ is the running variable, taking negative values before the transition date and positive values thereafter. β_1 measures the discontinuity at the transition date, the immediate impact of entering or leaving, respectively. As our baseline, we estimate $f(\cdot)$ linearly around the cutoff within a bandwidth of 30 days and using a triangular kernel.

Table 3.7 depicts the RD estimates. The estimated impacts on the time of the first and last action confirm the graphical evidence. In spring, the first and last action are delayed by 14 and 9 minutes, respectively, implying a slightly shorter work time overall (approximated as time of last action minus time of first action). By contrast, following the fall transition, the first and last action are conducted 13 and 10 minutes earlier, respectively. At the spring transition,

der a fixed social clock time causes a shift of between 9 and 26 minutes in average times of Twitter usage, commuting times and visits to businesses and other establishments in the US.

we find a reduction of 0.16 or 5.5% in total actions, while in fall actions increase by 0.09 or 3.3% of the sample mean. Factors that could drive these impacts on output include direct productivity effects of additional sleep (fall) or sleep loss (spring), and the changes in work time input described above. Moreover, when social clock times get detached from solar time in spring, circadian clocks can get disrupted giving rise to social jetlag (Kantermann et al., 2007; Roenneberg et al., 2019b). This disruption ends at the fall transition.

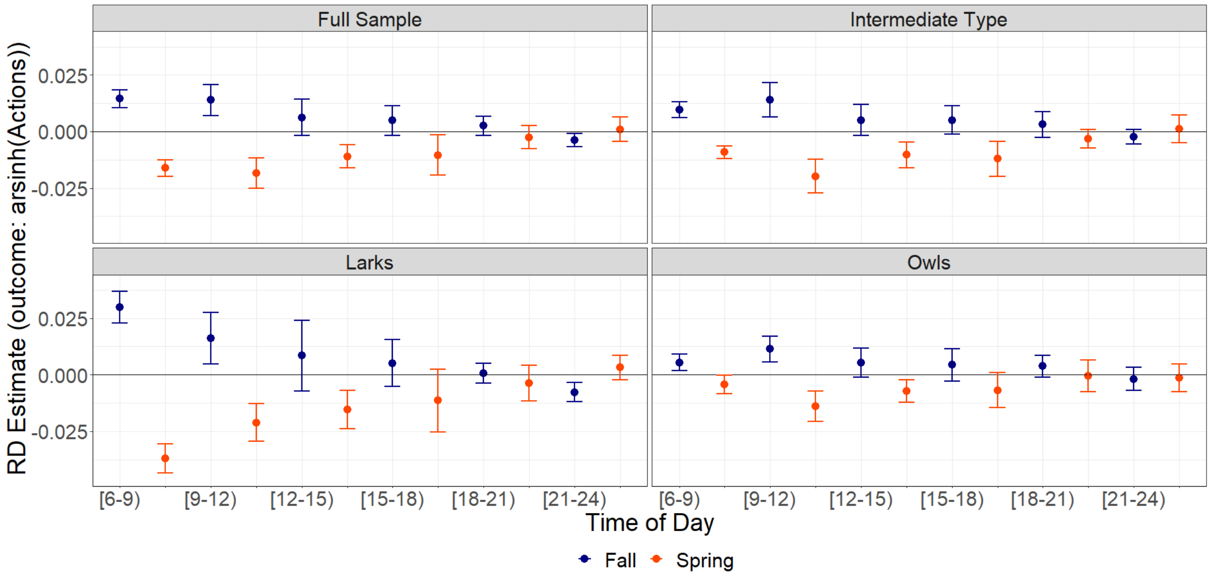
Table 3.7.: RD Estimates of Discontinuities at Transitions into and out of DST

Outcome		Full Sample	Larks	Intermed.	Owls
Actions	Fall	0.094	0.134	0.080	0.069
		(0.030)	(0.052)	(0.033)	(0.039)
		[3,608,156]	[1,077,285]	[1,448,138]	[1,082,733]
	Spring	-0.158	-0.253	-0.136	-0.091
		(0.036)	(0.050)	(0.036)	(0.038)
		[4,056,246]	[1,222,944]	[1,613,855]	[1,219,447]
arsinh(Actions)	Fall	0.020	0.030	0.016	0.014
		(0.008)	(0.013)	(0.007)	(0.006)
		[3,608,156]	[1,077,285]	[1,448,138]	[1,082,733]
	Spring	-0.028	-0.044	-0.025	-0.015
		(0.007)	(0.008)	(0.006)	(0.006)
		[4,056,246]	[1,222,944]	[1,613,855]	[1,219,447]
Time of First Action (minutes)	Fall	-13.39	-15.34	-12.80	-11.82
		(1.94)	(2.05)	(2.00)	(2.81)
		[1,368,770]	[450,428]	[531,757]	[386,585]
	Spring	14.29	18.76	12.86	10.92
		(2.29)	(3.22)	(2.77)	(2.20)
		[1,586,690]	[527,084]	[612,150]	[447,456]
Time of Last Action (minutes)	Fall	-9.94	-8.47	-10.00	-11.43
		(1.63)	(2.53)	(2.03)	(2.37)
		[1,368,770]	[450,428]	[531,757]	[386,585]
	Spring	9.25	7.05	8.84	12.33
		(2.72)	(3.87)	(2.13)	(3.07)
		[1,586,690]	[527,084]	[612,150]	[447,456]

Note: Table depicts estimates of the discontinuity in the variables listed on the left that occur right after the transitions into DST (Spring) and out of DST (Fall). All estimates are obtained from separate regressions, based on Equation 3.9. Discontinuities are estimated for the full sample (left) and separately by chronotype. Bandwidth = 30. Robust standard errors clustered at the developer level in parentheses. Effective observations in squared brackets.

We investigate effect heterogeneity by chronotype, and find that the changes in output at both transitions are largest for larks, and gradually shrink for later chronotypes. This holds both in absolute and relative terms. The shifts in work start time are also largest among larks at both transitions, while the effects on the time of the last action are smallest, which implies that total work time decreases most for larks after the spring transition (by approximately 9 minutes), and increases most after the fall transition (by 7 minutes). Among owls, by contrast, changes in time of first and last action are very similar both in fall and spring, implying minimal changes in workday length. Still, even in this group without changes in time input, we find a marginally significant increase in output in fall, and a significant decrease at the spring transition.

Figure 3.8.: RD Estimates for Different Times of the Day



Note: The plots depict estimates of the discontinuity in arsinh(actions) that occur right after the spring transitions into DST (blue) and the fall transition out of DST (red), along with 95% confidence intervals. Different estimates within one plot refer to actions during different three-hour windows of the day, as depicted on the x-axis. Discontinuities are estimated for the full sample and separately by chronotype. Upper left: full sample; upper right: intermediate chronotype; bottom left: larks; bottom right: owls. Bandwidth = 30.

Following Smith (2016), we also investigate effect heterogeneity across times of day. Figure 3.8 depicts the RD estimates where the outcomes are the inverse hyperbolic sine transformation of the number of actions conducted during several three-hour-windows of the day, in the full sample and separately by chronotype. At the fall transition (blue), we see increases in output in the morning, point estimates close to zero around lunch time and afternoon, and small decreases in the evening hours, consistent with the result that developers start and end work earlier. At the spring transition (red), the overall negative effect is mostly driven by reductions in activity in the morning, in line with later work start times, but effects also tend to be negative in the afternoon and only attenuate in the evening. Even though we found that the end of the work day is delayed, there are no positive effects in the evening, possibly due to lower productivity because of sleep deprivation. These patterns across hours of the day likely explain, why larks, who are particularly productive in the morning, are most affected. The impacts of the transition are strongest during their most productive time of day. Owls, who are less active and less productive during these hours are consequently less affected overall.

Moreover, adjusting the circadian clock to the new social clock after the spring transition is especially difficult for individuals who woke up before sunrise before the transition, and again wake up during darkness after the clock change. This might also contribute to the strong effects among larks who are most likely to be affected by this.

To check robustness of our main RD results, Appendix Figure 3.A.6 shows how the RD estimates vary as we change the bandwidth, from 20 to 60 in steps of 10. Results are overall robust to changing the bandwidth, but as the bandwidth is increased, they shrink somewhat in magnitude. Only the results on the change in actions at the fall transition are not consistently significant as the bandwidth is changed, but point estimates are mostly positive. Thus, we note that, while the spring transition clearly causes a decline in output, the increase in actions due to the fall transition is less clear. In Appendix Table 3.B.2 we conduct a placebo test. We repeat the RD analysis, but we shift the transition dates by 3 weeks into the future. Estimated effects are substantially smaller than at the true transition dates.

3.6. Conclusion

This paper provides evidence that a misalignment of the circadian clock and social schedules adversely affects knowledge worker performance. We conduct a causal analysis exploiting the transition into and out of daylight saving time as natural experiments. We find that the spring transition, which shortens nighttime and likely also sleep duration and leads to a deviation of social clocks from solar clocks, causes a reduction in daily output by 5.5%. We find positive but smaller effects for the fall transition. We complement this by showing that developers' performance also declines when they collaborate with users in other time zones and thus adapt their work schedules to times that are not in line with their natural rhythms. We also present a descriptive comparison of performance between developers with different time preferences, or chronotypes. We show that larks outperform later chronotypes on working days.

The DST transitions allow us to estimate causal effects and show that disruptions to circadian rhythms are indeed relevant in this setting. The overall economic costs of the short-run impacts are, however, relatively small.[24] The results on the impacts of collaboration across time zones imply that there are additional, potentially much larger costs due to misalignments among workers in remote teams, especially as remote work and cross-time zone collaboration are becoming increasingly common among knowledge workers. Finally, the more descriptive result on performance gaps between chronotypes suggests that, even in flexible and modern work environments, later chronotypes suffer from social jetlag. Given that we find that owls

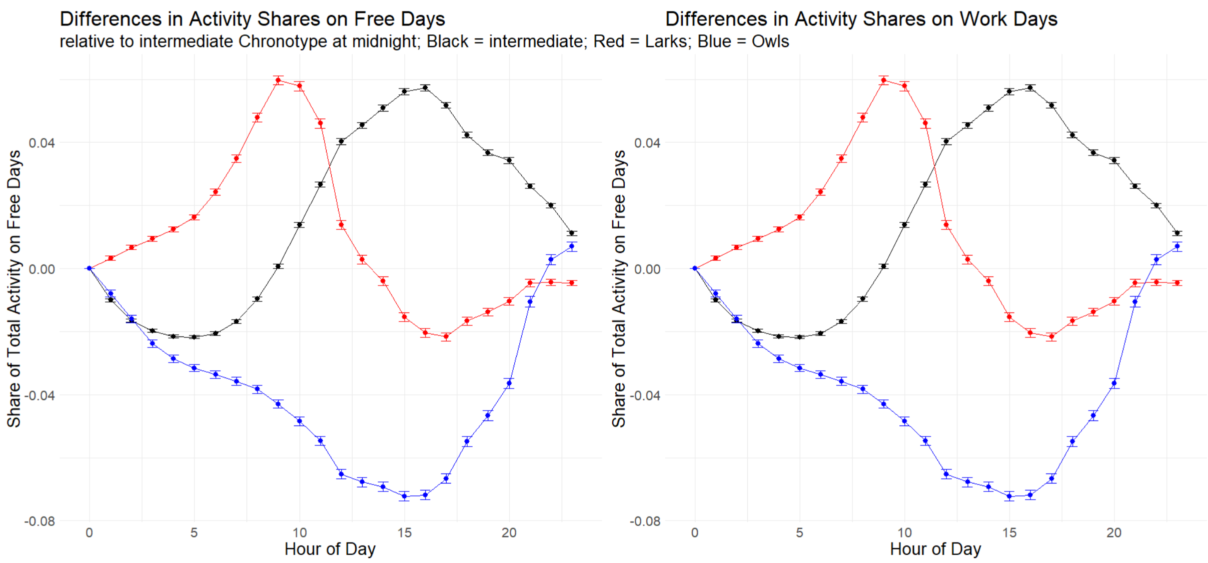
[24] This statement refers to cost through lost productivity analysed in this paper, not to results in other studies on consequences such as adverse health impacts or fatal traffic accidents.

make up 35% of our sample, this also hints at substantial costs and suggests that measures to better accommodate workers with different time preferences might be beneficial for worker and firm performance.

Appendix to Chapter 3

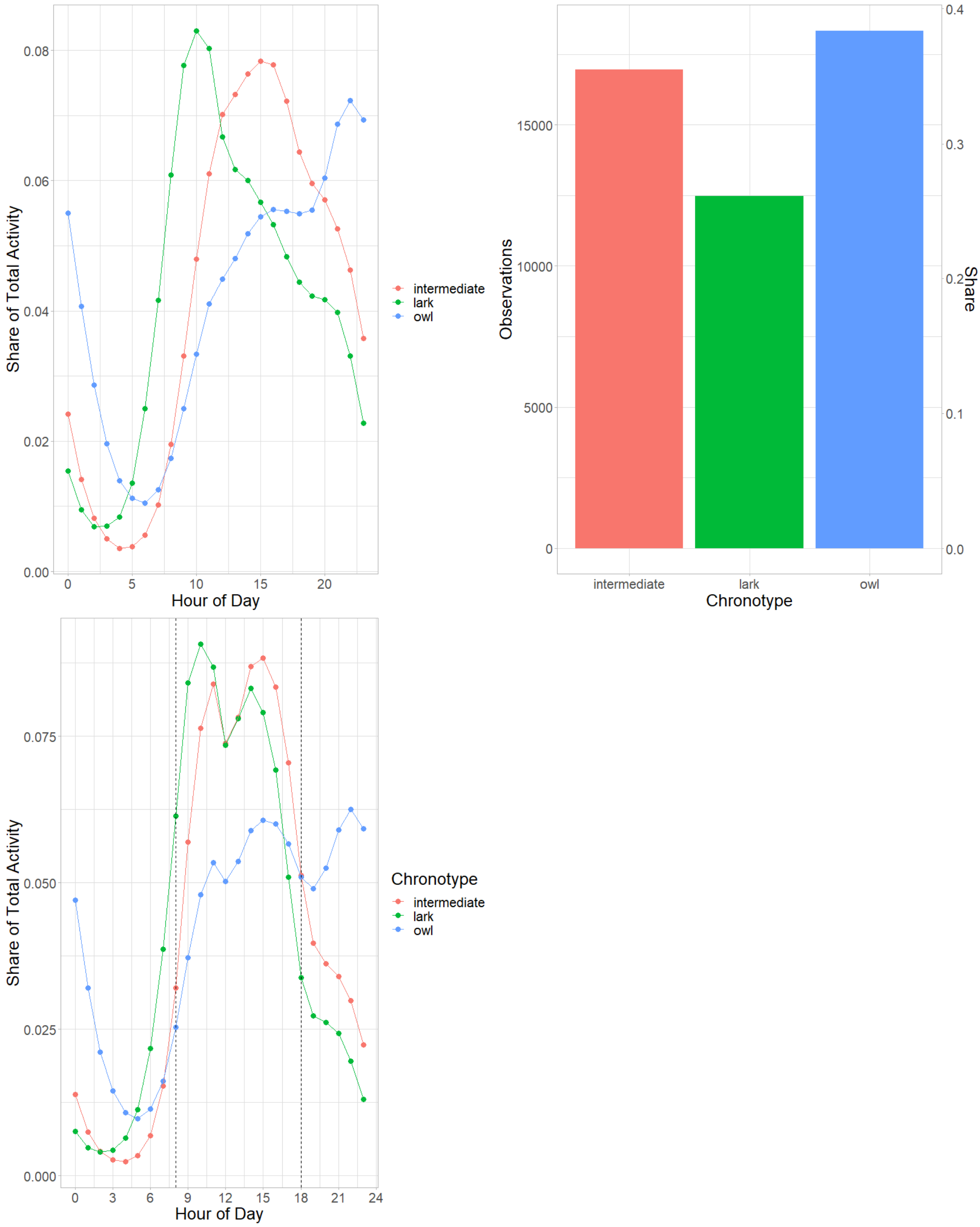
3.A. Additional Figures

Figure 3.A.1.: Differences in Activity Patterns across Chronotypes



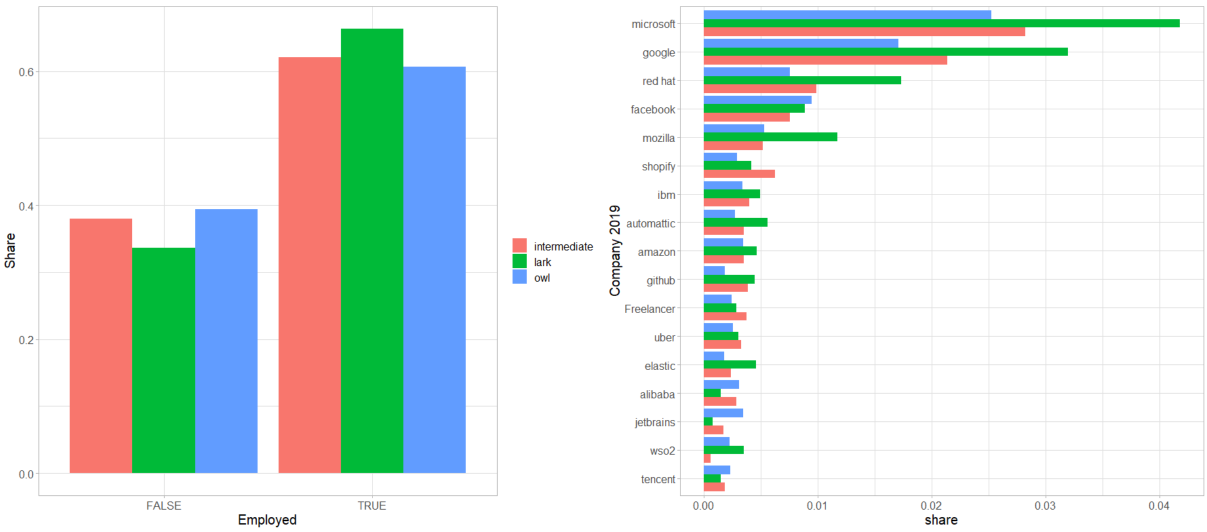
Notes: Plots show coefficients from regressions of hourly activity shares on 23 dummies for hours of the day (ommited hour is 0), and interactions of the hour-dummies with indicators for chronotype (intermediate type is omitted), along with 95% confidence intervalls.

Figure 3.A.2.: Chronotype Classification for Full Developer Sample



Notes: Upper left plot: Activity Patterns on weekends and public holidays by three types resulting from k-means clustering algorithm using the full developer sample. Upper right plot: Number of developers and share of total sample by k-means type. Bottom left plot: Activity Patterns on working days by three types resulting from k-means clustering algorithm. K-means clustering based on 16 variables measuring share of activities during eight three-hour windows of the day, separately for work days and free days.

Figure 3.A.3.: Employment Status by Chronotype



Notes: Left plot: Share of users reporting a company affiliation on their profile at the end of the sample period, by chronotype. Right plot: Share of users reporting to work at each of the 17 most frequent companies, by chronotype.

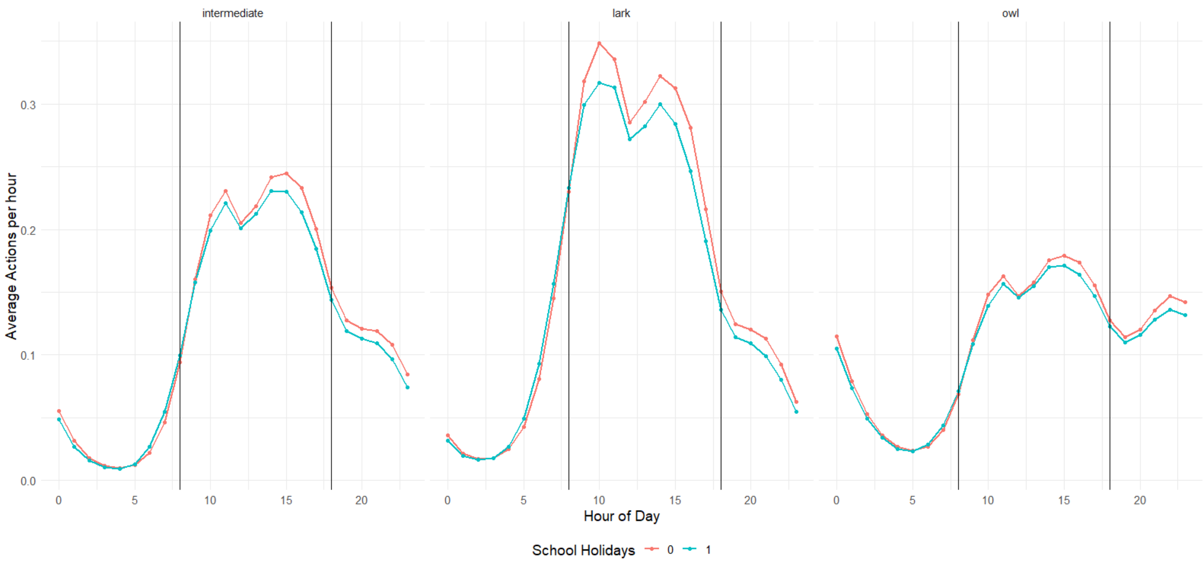
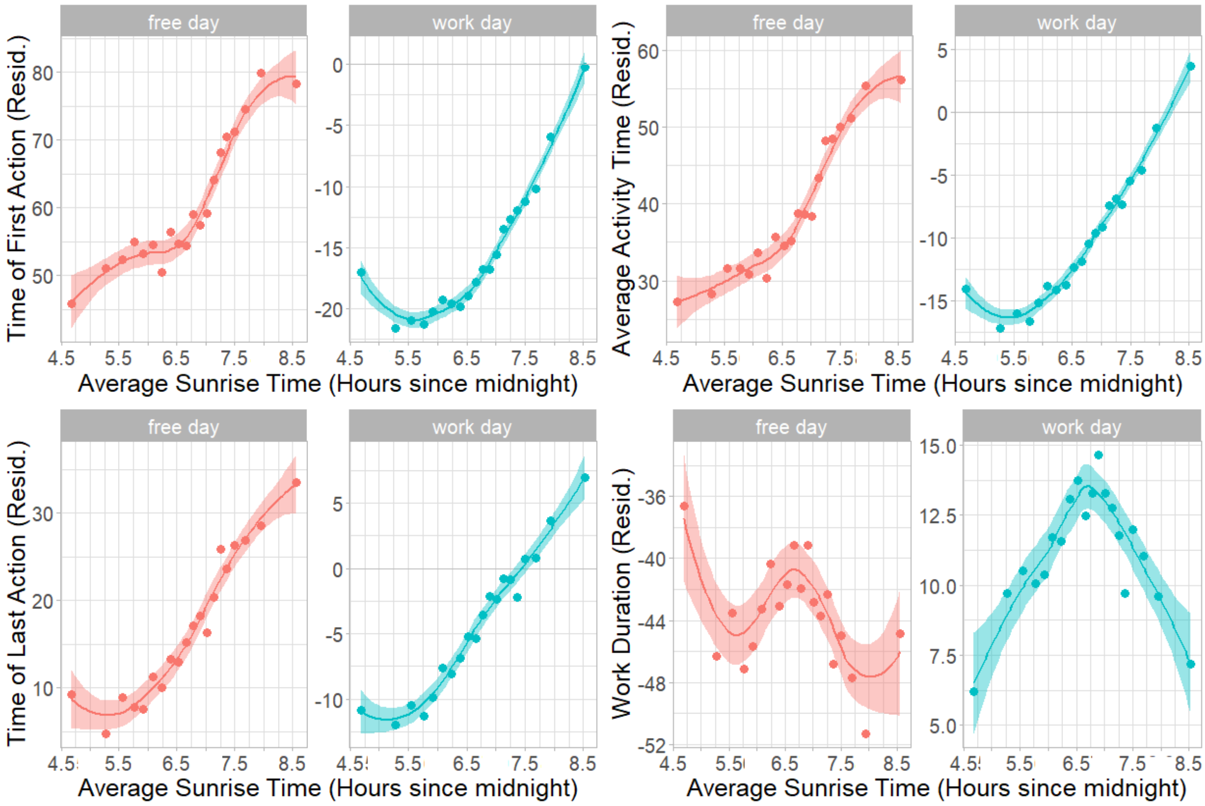


Figure 3.A.4.: School Holidays

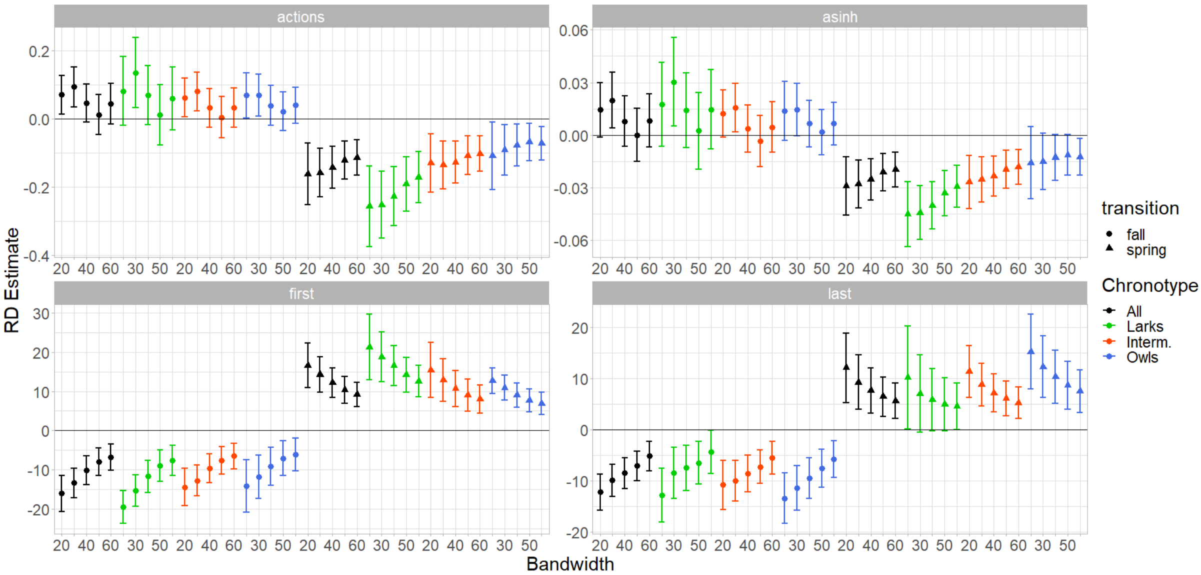
Notes: Lines depict mean number of actions per hour and day, averaged across all working days for users of the respective chronotype. Red line: Working day outside of school holidays. Blue line: Working days during school holidays.

Figure 3.A.5.: Work Time and Sunrise Time



Notes: Binned scatter plots depicting relationship between four variables assessing work time (in minutes) and sunrise time (in hours). Variables are minute of day of first action (upper left plots), minute of day of average action (upper right plots), minute of day of last action (bottom left plots), and work day length, approximated as difference between time of last and time of first action (bottom right plots). All four variables are residualized w.r.t. developer fixed effects. Relationship are depicted separately for working days (blue) and free days (red).

Figure 3.A.6.: RD Estimates for Different Bandwidth Choices



Notes: Figure depicts RD estimates with 95% confidence interval analogues to Table 3.7. For each outcome and estimation sample, five estimates are presented for five different bandwidth values, $\{20, 30, 40, 50, 60\}$. x-axis: bandwidth, color: estimation sample, shape: transition. Different plots refer to different outcome variables (Total actions, arsinh of total actions, minute of day of first action and minute of day of last action).

3.B. Additional Tables

Table 3.B.1.: Activity Differences between Chronotypes

	Actions			
Lark	0.820***	0.725***	0.627***	0.776***
	(0.099)	(0.091)	(0.078)	(0.071)
Owl	−0.206**	−0.326***	−0.029	−0.045
	(0.089)	(0.088)	(0.072)	(0.060)
Sample	Active Days	US only	Active Half	Company
Observations	10,768,932	10,381,814	16,499,506	19,448,401

Notes: Table displays estimated coefficients from Equation 3.4 estimated on different subsamples, as described at the bottom. Standard errors clustered at the developer level are reported in parentheses.

Table 3.B.2.: Placebo RD Estimates

	Actions	arsinh(Actions)	Time of First Action	Time of Last Action
Fall	-0.030	-0.002	0.357	1.574
	(.015)	(.002)	(.998)	(1.050)
Spring	-0.024	-0.006	-1.194	1.027
	(.014)	(.002)	(.928)	(.972)

Notes: Dates of fall and spring transition are shifted by 21 days into the future relative to the true transition dates. Bandwidth = 30. Robust standard errors in parentheses.

Bibliography

Adhvaryu, A., J. Fenske, and A. Nyshadham (2019): “Early Life Circumstance and Adult Mental Health,” *Journal of Political Economy*, 127, 1516–1549.

Adhvaryu, A., N. Kala, and A. Nyshadham (2018): “The Skills to Pay the Bills: Returns to on-the-job Soft Skills Training,” NBER Working Paper, National Bureau of Economic Research.

——— (2022): “Management and Shocks to Worker Productivity,” *Journal of Political Economy*, 130, 1–47.

Aguilar-Gomez, S., H. Dwyer, J. S. Graff Zivin, and M. J. Neidell (2022): “This is Air: The ”Non-Health” Effects of Air Pollution,” Working Paper 29848, National Bureau of Economic Research.

Ai, H., J. Wu, and Z. Zhou (2023): “The long-run effects of fetal PM2. 5 exposure on mental health: evidence from China,” *Environmental Science and Pollution Research*, 30, 34158–34173.

Akee, R., W. Copeland, E. J. Costello, and E. Simeonova (2018): “How Does Household Income Affect Child Personality Traits and Behaviors?” *American Economic Review*, 108, 775–827.

Alan, S., T. Boneva, and S. Ertac (2019): “Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit,” *The Quarterly Journal of Economics*, 134, 1121–1162.

Allen, T. A. and C. G. DeYoung (2017): “Personality Neuroscience and the Five Factor Model,” *Oxford Handbook of the Five Factor Model*, 319–352.

Almeida, S., M. Manousakas, E. Diapouli, Z. Kertesz, L. Samek, E. Hristova, K. Šega, R. P. Alvarez, C. Belis, and K. Eleftheriadis (2020): “Ambient particulate matter source apportionment using receptor modelling in European and Central Asia urban areas,” *Environmental Pollution*, 266, 115199.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011): “Personality Psychology and Economics,” in *Handbook of the Economics of Education*, Elsevier, vol. 4, 1–181.

Anderson, J. O., J. G. Thundiyil, and A. I. Stolbach (2011): “Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health,” *Journal of Medical Toxicology*, 8, 166–175.

Angelici, M. and P. Profeta (2020): "Smart-working: Work flexibility without constraints," Working paper, CESifo.

Annavarapu, R. N. and S. Kathi (2016): "Cognitive Disorders in Children Associated with Urban Vehicular Emissions," *Environmental Pollution*, 208, 74–78, special Issue: Urban Health and Wellbeing.

Arceo, E., R. Hanna, and P. Oliva (2016): "Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City," *The Economic Journal*, 126, 257–280.

Archsmith, J., A. Heyes, and S. Saberian (2018): "Air Quality and Error Quantity: Pollution and Performance in a High-Skilled, Quality-Focused Occupation," *Journal of the Association of Environmental and Resource Economists*, 5, 827–863.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The skill content of recent technological change: An empirical exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Autor, D. H. and B. Price (2013): "The Changing Task Composition of the US Labor Market: An Update of Autor, Levy, and Murnane (2003)," Working paper.

Babadjouni, R. M., D. M. Hodis, R. Radwanski, R. Durazo, A. Patel, Q. Liu, and W. J. Mack (2017): "Clinical effects of air pollution on the central nervous system; a review," *Journal of Clinical Neuroscience*, 43, 16–24.

Banzhaf, H. S., L. Ma, and C. Timmins (2019): "Environmental Justice: Establishing Causal Relationships," *Annual Review of Resource Economics*, 11, 377–398.

Barnes, C. M. and D. T. Wagner (2009): "Changing to daylight saving time cuts into sleep and increases workplace injuries." *Journal of applied psychology*, 94, 1305.

Barwick, P. J., S. Li, D. Rao, and N. Zahur (2018): "The morbidity cost of air pollution: evidence from consumer spending in China," Working Paper 2999068, SSRN.

Bassi, V., M. E. Kahn, N. Lozano Gracia, T. Porzio, and J. Sorin (2021): "Pollution in Ugandan Cities: Do Managers Avoid it or Adapt in Place?" Working Paper 3887079, SSRN.

Baylis, P. (2020): "Temperature and temperament: Evidence from Twitter," *Journal of Public Economics*, 184, 104161.

Baylis, P., S. Borenstein, and E. A. Rubin (2023): "When we change the clock, does the clock change us?" Tech. rep., National Bureau of Economic Research.

Beckmann, M., T. Cornelissen, and M. Kräkel (2017): "Self-managed working time and employee effort: Theory and evidence," *Journal of Economic Behavior & Organization*, 133, 285–302.

Beckwith, T., K. Cecil, M. Altaye, R. Severs, C. Wolfe, Z. Percy, T. Maloney, K. Yolton, G. LeMasters, K. Brunst, and P. Ryan (2020): "Reduced Gray Matter Volume and Cortical Thickness Associated with Traffic-related Air Pollution in a Longitudinally Studied Pediatric

Cohort," *PLOS ONE*, 15, 1–19.

Bharadwaj, P., M. Gibson, J. G. Zivin, and C. Neilson (2017): "Gray matters: Fetal Pollution Exposure and Human Capital Formation," *Journal of the Association of Environmental and Resource Economists*, 4, 505–542.

Bishop, K. C., J. D. Ketcham, and N. V. Kuminoff (forthcoming): "Hazed and confused: the effect of air pollution on dementia," *The Review of Economic Studies*.

Black, S. E., E. Grönqvist, and B. Öckert (2018): "Born to lead? The Effect of Birth Order on Noncognitive Abilities," *Review of Economics and Statistics*, 100, 274–286.

Bonke, J. (2012): "Do morning-type people earn more than evening-type people? How chronotypes influence income," *Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE*, 55–72.

Borghans, L., H. Meijers, and B. Ter Weel (2008): "The Role of Noncognitive Skills in Explaining Cognitive Test Scores," *Economic Inquiry*, 46, 2–12.

Borgschulte, M., D. Molitor, and E. Zou (forthcoming): "Air Pollution and the Labor Market: Evidence from Wildfire Smoke," *The Review of Economics and Statistics*.

Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (2002): "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence," *The Quarterly Journal of Economics*, 117, 339–376.

Bureau of Labor Statistics (2021): "Software Developers, Quality Assurance Analysts, and Testers," in *Occupational Outlook Handbook*, U.S. Department of Labor.

——— (2022): "Occupational Employment and Wage Statistics," https://www.bls.gov/oes/current/oes_nat.htm, accessed: 2022-11-07.

Burgess, H. J., F. Kikyo, Z. Valdespino-Hayden, M. Rizvydeen, M. Kimura, M. H. Pollack, S. E. Hobfoll, K. B. Rajan, A. K. Zalta, and J. W. Burns (2018): "Do the Morningness-Eveningness questionnaire and Munich ChronoType questionnaire change after morning light treatment?" *Sleep science and practice*, 2, 1–5.

Burke, M., S. Heft-Neal, J. Li, A. Driscoll, P. W. Baylis, M. Stigler, J. Weill, J. Burney, J. Wen, M. Childs, and C. Gould (2022): "Exposures and Behavioral Responses to Wildfire Smoke," *Nature Human Behavior*.

Calderón-Garcidueñas, L., M. Franco-Lira, R. Torres-Jardón, C. Henriquez-Roldán, G. Barragán-Mejía, G. Valencia-Salazar, A. Gonzaléz-Maciel, R. Reynoso-Robles, R. Villarreal-Calderón, and W. Reed (2007): "Pediatric Respiratory and Systemic Effects of Chronic Air Pollution Exposure: Nose, Lung, Heart, and Brain Pathology," *Toxicologic Pathology*, 35, 154–162.

Carneiro, P., C. Crawford, and A. Goodman (2007): "The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes," CEE DP 92.

Cedefop (2022): "Employed population by occupation and sector," https://www.cedefop.europa.eu/en/tools/skills-intelligence/employed-population-occupation-and-sector?year=2020&country=EU&occupation=#1, accessed: 2022-11-07.

Chang, T., J. Graff Zivin, T. Gross, and M. Neidell (2016): "Particulate Pollution and the Productivity of Pear Packers," *American Economic Journal: Economic Policy*, 8, 141–169.

Chang, T. Y., J. Graff Zivin, T. Gross, and M. Neidell (2019): "The Effect of Pollution on Worker Productivity: Evidence from Call Center Workers in China," *American Economic Journal: Applied Economics*, 11, 151–172.

Chauvin, J., P. Choudhury, and T. P. Fang (2021): "The effects of temporal distance on intra-firm communication: Evidence from daylight savings time," *Harvard Business School Technology & Operations Mgt. Unit Working Paper*.

Cicero-Fernandez, P., W. Thistlewaite, Y. Falcon, and I. Guzman (1993): "TSP, PM10 and PM10/TSP ratios in the Mexico City Metropolitan area: a temporal and spatial approach." *Journal of exposure analysis and environmental epidemiology*, 3, 1–14.

Colmer, J., D. Lin, S. Liu, and J. Shimshack (2021a): "Why are pollution damages lower in developed countries? Insights from high-Income, high-particulate matter Hong Kong," *Journal of Health Economics*, 79, 102511.

——— (2021b): "Why are Pollution Damages Lower in Developed Countries? Insights from High-Income, High-Particulate Matter Hong Kong," *Journal of Health Economics*, 102511.

Colmer, J. and J. Voorheis (2020): "The Grandkids aren't alright: The Intergenerational Effects of Prenatal Pollution Exposure," CEP Discussion Papers 1733, Centre for Economic Performance, LSE.

Conlin, A., I. Nerg, L. Ala-Mursula, T. Räihä, and M. Korhonen (2022): "Chronotype and Wages at Mid-Age," *Available at SSRN 4240603*.

Costa, L. G., T. B. Cole, J. Coburn, Y.-C. Chang, K. Dao, and P. Roque (2014): "Neurotoxicants are in the Air: Convergence of Human, Animal, and in vitro Studies on the Effects of Air Pollution on the Brain," *BioMed Research International*, 2014.

Costa-i Font, J., S. Fleche, and R. Pagan (2022): "The Labour Market Returns to Sleep," .

Cubel, M., A. Nuevo-Chiquero, S. Sanchez-Pages, and M. Vidal-Fernandez (2016): "Do Personality Traits Affect Productivity? Evidence from the Laboratory," *The Economic Journal*, 126, 654–681.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Currie, J. and D. Almond (2011): "Chapter 15 - Human Capital Development Before Age Five," Elsevier, vol. 4 of *Handbook of Labor Economics*, 1315–1486.

Currie, J., L. Davis, M. Greenstone, and R. Walker (2015): "Environmental Health Risks and Housing Values: Evidence from 1,600 Toxic Plant Openings and Closings," *American Economic Review*, 105, 678–709.

Currie, J. and M. Neidell (2005): "Air Pollution and Infant Health: What Can We Learn from California's Recent Experience?" *The Quarterly Journal of Economics*, 120, 1003–1030.

Currie, J. and M. Stabile (2006): ""Child Mental Health and Human Capital Accumulation: The Case of ADHD"," *Journal of Health Economics*, 25, 1094–1118.

Currie, J., J. Voorheis, and R. Walker (2023): "What caused racial disparities in particulate exposure to fall? New evidence from the Clean Air Act and satellite-based measures of air quality," *American Economic Review*, 113, 71–97.

Datta Gupta, N. and M. Simonsen (2010): "Non-cognitive Child Outcomes and Universal High Quality Child Care," *Journal of Public Economics*, 94, 30–43.

De Pauw, S. S. (2017): "Childhood Personality and Temperament," *The Oxford Handbook of the Five Factor Model*, 243–280.

de Prado Bert, P., E. M. H. Mercader, J. Pujol, J. Sunyer, and M. Mortamais (2018): "The Effects of Air Pollution on the Brain: A Review of Studies Interfacing Environmental Epidemiology and Neuroimaging," *Current environmental health reports*, 5, 351–364.

Deal, J. E., C. F. Halverson Jr, V. Havill, and R. P. Martin (2005): "Temperament Factors as Longitudinal Predictors of Young Adult Personality," *Merrill-Palmer Quarterly (1982-)*, 315–334.

Delgado-Saborit, J. M., V. Guercio, A. M. Gowers, G. Shaddick, N. C. Fox, and S. Love (2021): "A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population," *Science of The Total Environment*, 757, 143734.

Deming, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 132, 1593–1640.

Deng, G., Z. Li, Z. Wang, J. Gao, Z. Xu, J. Li, and Z. Wang (2017): "Indoor/outdoor relationship of PM2.5 concentration in typical buildings with and without air cleaning in Beijing," *Indoor and Built Environment*, 26, 60–68.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019): "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 109, 4178–4219.

Donald, S. G. and K. Lang (2007): "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, 89, 221–233.

Ebenstein, A., V. Lavy, and S. Roth (2016): "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution," *American Economic Journal: Applied Economics*, 8, 36–65.

EDIN, P.-A., P. FREDRIKSSON, M. NYBOM, AND B. ÖCKERT (2017): "The Rising Return to Non-Cognitive Skill," Discussion paper, IZA.

FALK, A., T. NEUBER, AND P. STRACK (2021): "Limited Self-Knowledge and Survey Response Behavior," Discussion Paper 307, CRC TR 224.

FISCHER, D., D. A. LOMBARDI, H. MARUCCI-WELLMAN, AND T. ROENNEBERG (2017): "Chronotypes in the US–influence of age and sex," *PloS one*, 12, e0178782.

FLETCHER, J. M. (2013): "The Effects of Personality Traits on Adult Labor Market Outcomes: Evidence from Siblings," *Journal of Economic Behavior & Organization*, 89, 122–135.

FU, S., V. B. VIARD, AND P. ZHANG (2021): "Air Pollution and Manufacturing Firm Productivity: Nationwide Estimates for China," *The Economic Journal*, 131, 3241–3273.

GAGGERO, A. AND D. TOMMASI (2022): "Time of Day and High Stakes Cognitive Assessments," *The Economic Journal*.

GIBSON, M. AND J. SHRADER (2018): "Time use and labor productivity: The returns to sleep," *Review of Economics and Statistics*, 100, 783–798.

GIUNTELLA, O., W. HAN, AND F. MAZZONNA (2017): "Circadian rhythms, sleep, and cognitive skills: Evidence from an unsleeping giant," *Demography*, 54, 1715–1742.

GIUNTELLA, O. AND F. MAZZONNA (2019): "Sunset time and the economic effects of social jetlag: evidence from US time zone borders," *Journal of health economics*, 65, 210–226.

GLATTER-GÖTZ, H., P. MOHAI, W. HAAS, AND C. PLUTZAR (2019): "Environmental Inequality in Austria: Do Inhabitants' Socioeconomic Characteristics Differ Depending on their Proximity to Industrial Polluters?" *Environmental Research Letters*, 14, 074007.

GODZINSKI, A. AND M. S. CASTILLO (2021): "Disentangling the Effects of Air Pollutants with many Instruments," *Journal of Environmental Economics and Management*, 109, 102489.

GOEBEL, J., M. M. GRABKA, S. LIEBIG, M. KROH, D. RICHTER, C. SCHRÖDER, AND J. SCHUPP (2019): "The German socio-economic panel (SOEP)," *Jahrbücher für Nationalökonomie und Statistik*, 239, 345–360.

GOLDIN, A. P., M. SIGMAN, G. BRAIER, D. A. GOLOMBEK, AND M. J. LEONE (2020): "Interplay of chronotype and school timing predicts school performance," *Nature Human Behaviour*, 4, 387–396.

GRAFF ZIVIN, J. AND M. NEIDELL (2012): "The Impact of Pollution on Worker Productivity," *American Economic Review*, 102, 3652–3673.

——— (2014): "Temperature and the Allocation of Time: Implications for Climate Change," *Journal of Labor Economics*, 32, 1–26.

GRÖNQVIST, E., B. ÖCKERT, AND J. VLACHOS (2017): "The Intergenerational Transmission of Cognitive and Noncognitive Abilities," *Journal of Human Resources*, 52, 887–918.

GRÖNQVIST, H., J. P. NILSSON, AND P.-O. ROBLING (2020): "Understanding how low levels of early lead exposure affect children's life trajectories," *Journal of Political Economy*, 128, 3376–3433.

Hafner, M., M. Stepanek, J. Taylor, W. M. Troxel, and C. V. Stolk (2016): *Why sleep matters: the economic costs of insufficient sleep: A cross-country comparative analysis*, Santa Monica, CA: RAND Corporation.

Hamermesh, D. S., C. K. Myers, and M. L. Pocock (2008): "Cues for timing and coordination: latitude, letterman, and longitude," *Journal of Labor Economics*, 26, 223–246.

He, J., H. Liu, and A. Salvo (2019): "Severe Air Pollution and Labor Productivity: Evidence from Industrial Towns in China," *American Economic Journal: Applied Economics*.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411–482.

Heyes, A., N. Rivers, and B. Schaufele (2019): "Pollution and Politician Productivity: The Effect of PM on MPs," *Land Economics*, 95, 157–173.

Hoek, G., G. Kos, R. Harrison, J. de Hartog, K. Meliefste, H. ten Brink, K. Katsouyanni, A. Karakatsani, M. Lianou, A. Kotronarou, I. Kavouras, J. Pekkanen, M. Vallius, M. Kulmala, A. Puustinen, S. Thomas, C. Meddings, J. Ayres, J. van Wijnen, and K. Hameri (2008): "Indoor–outdoor relationships of particle number and mass in four European cities," *Atmospheric Environment*, 42, 156–169.

Hoffmann, B. and J. P. Rud (2022): "Exposure or Income? The Unequal Effects of Pollution on Daily Labor Supply," Working paper, Inter-American Development Bank.

Horne, J. A. and O. Östberg (1976): "A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms." *International journal of chronobiology*.

Hsiang, S., P. Oliva, and R. Walker (2019): "The Distribution of Environmental Damages," *Review of Environmental Economics and Policy*, 13, 83–103.

Huang, J., N. Xu, and H. Yu (2020): "Pollution and Performance: Do Investors Make Worse Trades on Hazy Days?" *Management Science*, 66, 4455–4476.

Hufe, P. (2020): "The Parental Wage Gap and the Development of Socio-emotional Skills in Children," .

Isen, A., M. Rossin-Slater, and W. R. Walker (2017a): "Every Breath You Take—Every Dollar You'll Make: The Long-Term Consequences of the Clean Air Act of 1970," *Journal of Political Economy*, 125, 848–902.

Isen, A., M. Rossin-Slater, and W. R. Walker (2017b): "Every Breath You Take—Every Dollar You'll Make: The Long-term Consequences of the Clean Air Act of 1970," *Journal of Political Economy*, 125, 848–902.

JAGNANI, M. (2022): "Children's Sleep and Human Capital Production," *The Review of Economics and Statistics*, 1–45.

JANS, J., P. JOHANSSON, AND J. P. NILSSON (2018a): "Economic status, air quality, and child health: Evidence from inversion episodes," *Journal of Health Economics*, 61, 220–232.

——— (2018b): "Economic Status, Air Quality, and Child Health: Evidence from Inversion Episodes," *Journal of Health Economics*, 61, 220–232.

JIN, L. AND N. R. ZIEBARTH (2020): "Sleep, health, and human capital: Evidence from daylight saving time," *Journal of Economic Behavior & Organization*, 170, 174–192.

JOHNSON, N. M., A. R. HOFFMANN, J. C. BEHLEN, C. LAU, D. PENDLETON, N. HARVEY, R. SHORE, Y. LI, J. CHEN, Y. TIAN, ET AL. (2021): "Air pollution and Children's Health—A Review of Adverse Effects Associated with Prenatal Exposure from Fine to Ultrafine Particulate Matter," *Environmental Health and Preventive Medicine*, 26, 1–29.

JOHNSTON, D., C. PROPPER, S. PUDNEY, AND M. SHIELDS (2014): "Child Mental Health and Educational Attainment: Multiple Observers and the Measurement Error Problem," *Journal of Applied Econometrics*, 29, 880–900.

JONES, S. E., J. TYRRELL, A. R. WOOD, R. N. BEAUMONT, K. S. RUTH, M. A. TUKE, H. YAGHOOTKAR, Y. HU, M. TEDER-LAVING, C. HAYWARD, ET AL. (2016): "Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci," *PLoS genetics*, 12, e1006125.

KAHN, M. E. AND P. LI (2020): "Air pollution lowers high skill public sector worker productivity in China," *Environmental Research Letters*, 15, 084003.

KANTERMANN, T., M. JUDA, M. MERROW, AND T. ROENNEBERG (2007): "The human circadian clock's seasonal adjustment is disrupted by daylight saving time," *Current Biology*, 17, 1996–2000.

KARAGULIAN, F., C. A. BELIS, C. F. C. DORA, A. M. PRÜSS-USTÜN, S. BONJOUR, H. ADAIR-ROHANI, AND M. AMANN (2015): "Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level," *Atmospheric Environment*, 120, 475–483.

KAUR, S., S. MULLAINATHAN, S. OH, AND F. SCHILBACH (2021): "Do Financial Concerns Make Workers Less Productive?" Working Paper 28338, National Bureau of Economic Research.

KELLY, F. J. AND J. C. FUSSELL (2015): "Air pollution and public health: emerging hazards and improved understanding of risk," *Environmental Geochemistry and Health*, 37, 631 – 649.

KLAUBER, H., F. HOLUB, N. KOCH, N. PESTEL, N. RITTER, AND A. ROHLF (2021): "Killing Prescriptions Softly: Low Emission Zones and Child Health from Birth to School," Discussion paper, IZA.

KREBS, B. AND S. LUECHINGER (2021): "Air Pollution, Cognitive Performance, and the Role of

Task Proficiency," Working Paper 3947149, SSRN.

Künn, S., J. Palacios, and N. Pestel (forthcoming): "Indoor Air Quality and Strategic Decision-Making," *Management Science*.

La Nauze, A. and E. R. Severnini (2021): "Air Pollution and Adult Cognition: Evidence from Brain Training," Working Paper 28785, National Bureau of Economic Research.

Lai, S., A. Sorichetta, J. Steele, C. W. Ruktanonchai, A. D. Cunningham, G. Rogers, P. Koper, D. Woods, M. Bondarenko, N. W. Ruktanonchai, et al. (2022): "Global holiday datasets for understanding seasonal human mobility and population dynamics," *Scientific Data*, 9, 17.

Lazear, E. P., K. L. Shaw, and C. Stanton (2015): "Making Do with Less: Working Harder during Recessions," *Journal of Labor Economics*, 34, S333–S360.

Lederer, A. M., P. M. Fredriksen, B. N. Nkeh-Chungag, F. Everson, H. Strijdom, P. De Boever, and N. Goswami (2021): "Cardiovascular effects of air pollution: current evidence from animal and human studies," *American Journal of Physiology-Heart and Circulatory Physiology*, 320, H1417–H1439, pMID: 33513082.

Levy, R. J. (2015): "Carbon monoxide pollution and neurodevelopment: A public health concern," *Neurotoxicology and Teratology*, 49, 31–40.

Li, H., J. Cai, R. Chen, Z. Zhao, Z. Ying, L. Wang, J. Chen, K. Hao, P. L. Kinney, H. Chen, and H. Kan (2017): "Particulate Matter Exposure and Stress Hormone Levels: A Randomized, Double-blind, Crossover Trial of Air Purification," *Circulation*, 136, 618–627.

Lindqvist, E. and R. Vestman (2011): "The Labor Market Returns to Cognitive and Noncognitive ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3, 101–28.

LoPalo, M. (forthcoming): "Temperature, Worker Productivity, and Adaptation: Evidence from Survey Data Production," *American Economic Journal: Applied Economics*.

Mas, A. and A. Pallais (2020): "Alternative Work Arrangements," *Annual Review of Economics*, 12, 631–658.

McDermott, G. R. and B. Hansen (2021): "Labor Reallocation and Remote Work During COVID-19: Real-time Evidence from GitHub," Working Paper 29598, National Bureau of Economic Research.

Menon, S., A. Salvatori, and W. Zwysen (2020): "The effect of computer use on work discretion and work intensity: Evidence from Europe," *British Journal of Industrial Relations*, 58, 1004–1038.

Molina, T. (2021): "Pollution, Ability, and Gender-specific Investment Responses to Shocks," *Journal of the European Economic Association*, 19, 580–619.

Moutafi, J., A. Furnham, and I. Tsaousis (2006): "Is the Relationship between Intelligence and Trait Neuroticism Mediated by Test Anxiety?" *Personality and Individual Differences*, 40, 587–597.

Mullins, J. T. and C. White (2019): "Temperature and mental health: Evidence from the spectrum of mental health outcomes," *Journal of Health Economics*, 68, 102240.

Muro, M., S. Liu, J. Whiton, and S. Kulkarni (2017): "Digitalization and the American workforce," Working paper, Brookings India.

Myhre, O., M. Låg, G. D. Villanger, B. Oftedal, J. Øvrevik, J. A. Holme, H. Aase, R. E. Paulsen, A. Bal-Price, and H. Dirven (2018): "Early life Exposure to Air Pollution Particulate Matter (PM) as Risk Factor for Attention Deficit/Hyperactivity Disorder (ADHD): Need for Novel Strategies for Mechanisms and Causalities," *Toxicology and Applied Pharmacology*, 354, 196–214.

Neidell, M., J. G. Zivin, M. Sheahan, J. Willwerth, C. Fant, M. Sarofim, and J. Martinich (2021): "Temperature and Work: Time Allocated to Work under Varying Climate and Labor Market Conditions," *PLOS ONE*, 16, e0254224.

Papatsimpa, C., L. Schlangen, K. Smolders, J.-P. Linnartz, and Y. De Kort (2021): "The interindividual variability of sleep timing and circadian phase in humans is influenced by daytime and evening light conditions," *Scientific Reports*, 11, 13709.

Park, R. J. (2020): "Hot temperature and high stakes performance," *Journal of Human Resources*.

Pencavel, J. (2015): "The Productivity of Working Hours," *The Economic Journal*, 125, 2052–2076.

Persson, P. and M. Rossin-Slater (2018): "Family Ruptures, Stress, and the Mental Health of the Next Generation," *American Economic Review*, 108, 1214–52.

Pope, C. A. (2000): "Epidemiology of Fine Particulate Air Pollution and Human Health: Biologic Mechanisms and Who's at Risk?" *Environmental Health Perspectives*, 108, 713–723.

Pope, N. G. (2016): "How the time of day affects productivity: Evidence from school schedules," *Review of Economics and Statistics*, 98, 1–11.

Roenneberg, T. (2017): "Twitter as a means to study temporal behaviour," *Current Biology*, 27, R830–R832.

Roenneberg, T., C. J. Kumar, and M. Merrow (2007): "The human circadian clock entrains to sun time," *Current Biology*, 17, R44–R45.

Roenneberg, T., L. K. Pilz, G. Zerbini, and E. C. Winnebeck (2019a): "Chronotype and social jetlag: a (self-) critical review," *Biology*, 8, 54.

Roenneberg, T., E. C. Winnebeck, and E. B. Klerman (2019b): "Daylight saving time and artificial time zones–a battle between biological and social times," *Frontiers in physiology*, 10, 944.

Rosales-Rueda, M. and M. Triyana (2019): "The Persistent Effects of early-life Exposure to

Air Pollution: Evidence from the Indonesian Forest Fires," *Journal of Human Resources*, 54, 1037–1080.

Rüttenauer, T. (2018): "Neighbours Matter: A nation-wide small-area Assessment of Environmental Inequality in Germany," *Social Science Research*, 70, 198–211.

Sanders, N. J. (2012): "What Doesn't Kill You Makes You Weaker: Prenatal Pollution Exposure and Educational Outcomes," *Journal of Human Resources*, 47, 826–850.

Sarmiento, L. (2022): "Air pollution and the productivity of high-skill labor: evidence from court hearings," *The Scandinavian Journal of Economics*, 124, 301–332.

Shangguan, R., J. DeVaro, and O. Hideo (2021): "Enhancing Team Productivity through Shorter Working Hours: Evidence from the Great Recession," Working paper, The Research Institute of Economy, Trade and Industry.

Shepard, E. M. I., T. J. Clifton, and D. Kruse (1996): "Flexible work hours and productivity: Some evidence from the pharmaceutical industry," *Industrial Relations: A Journal of Economy and Society*, 35, 123–139.

Smarr, B. L. and A. E. Schirmer (2018): "3.4 million real-world learning management system logins reveal the majority of students experience social jet lag correlated with decreased performance," *Scientific Reports*, 8, 4793.

Smith, A. C. (2016): "Spring forward at your own risk: Daylight saving time and fatal vehicle crashes," *American Economic Journal: Applied Economics*, 8, 65–91.

SOEP (2019): "Socio-Economic Panel, data for years 1984-2018. DOI: 10.5684/soep-core.v35," .

Somanathan, E., R. Somanathan, A. Sudarshan, and M. Tewari (2021): "The impact of temperature on productivity and labor supply: Evidence from Indian manufacturing," *Journal of Political Economy*, 129, 1797–1827.

Sorrenti, G., U. Zölitz, D. Ribeaud, and M. Eisner (2020): "The Causal Impact of Socio-emotional Skills Training on Educational Success," Working Paper 343, University of Zurich, Department of Economics.

Tan, X., M. Zhou, and Z. Sun (2020): "A first look at good first issues on github," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 398–409.

Van der Meulen, A., B. Van Elzakker, and G. Van den Hooff (1987): "PM10: results of a one-year monitoring survey in the Netherlands," *JAPCA*, 37, 812–818.

Voorheis, J. (2017): "Air Quality, Human Capital Formation and the Long-term Effects of Environmental Inequality at Birth," Working paper, Center for Economic Studies, US Census Bureau.

Widiger, T. A., W. L. Gore, C. Crego, S. L. Rojas, and J. R. Oltmanns (2017): "Five Factor Model and Personality Disorder," *The Oxford Handbook of the Five Factor Model*, 449.

Woodward, N., A. Haghani, R. Johnson, T. Hsu, A. Saffari, C. Sioutas, S. Kanoski, C. Finch, and T. Morgan (2018): "Prenatal and early life Exposure to Air Pollution Induced Hippocampal Vascular Leakage and Impaired Neurogenesis in Association with Behavioral Deficits," *Translational Psychiatry*, 8, 1–10.

Xu, R., X. Qi, G. Dai, H. Lin, P. Zhai, C. Zhu, L. Wang, and A. Ding (2020): "A Comparison Study of Indoor and Outdoor Air Quality in Nanjing, China," *Aerosol and Air Quality Research*, 20, 2128–2141.

Xu, X., S. U. Ha, and R. Basnet (2016): "A Review of Epidemiological Research on Adverse Neurological Effects of Exposure to Ambient Air Pollution," *Frontiers in Public Health*, 4, 157.

Zerbini, G. and M. Merrow (2017): "Time to learn: H ow chronotype impacts education," *PsyCh journal*, 6, 263–276.

Zerbini, G., V. van der Vinne, L. K. Otto, T. Kantermann, W. P. Krijnen, T. Roenneberg, and M. Merrow (2017): "Lower school performance in late chronotypes: underlying factors and mechanisms," *Scientific reports*, 7, 1–10.

Curriculum Vitae

2017 - 2023	PhD in Economics *University of Mannheim*
2017 - 2019	M.Sc. in Economics *University of Mannheim*
2013 - 2017	B.Sc. in Economics *University of Cologne*