

NBER WORKING PAPER SERIES

SHIFTING WORK PATTERNS WITH GENERATIVE AI

Eleanor W. Dillon
Sonia Jaffe
Nicole Immorlica
Christopher T. Stanton

Working Paper 33795
<http://www.nber.org/papers/w33795>

NATIONAL BUREAU OF ECONOMIC RESEARCH

1050 Massachusetts Avenue
Cambridge, MA 02138
May 2025, Revised November 2025

We thank the Microsoft Customer Research Program, especially Alexia Cambon, Sida Peng, Neha Shah, Chris Gideon, Modern Work Marketing, the Office of Applied Research, and company partners for help carrying out this experiment, Abigail Atchison, Roman Basko, and Fabio Vera for superb data science support, Jack Cenatempo, Esther Plotnick, and Will Wang for excellent research assistance. We thank 3 anonymous referees, Tatyana Deryugina, Seema Jayachandran, Rem Koning, Anders Humlum, and Danielle Li for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at <http://www.nber.org/papers/w33795>

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2025 by Eleanor W. Dillon, Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Shifting Work Patterns with Generative AI

Eleanor W. Dillon, Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton

NBER Working Paper No. 33795

May 2025, Revised November 2025

JEL No. L23, M1, M15, M5, O33

ABSTRACT

We present evidence from a field experiment across 66 firms and 7,137 knowledge workers. Workers were randomly selected to access a generative AI tool integrated into applications they already used at work for email, meetings, and writing. In the second half of the 6-month experiment, the 80% of treated workers who used this tool spent two fewer hours on email each week and reduced their time working outside of regular hours. Apart from these individual time savings, we do not detect shifts in the quantity or composition of workers' tasks resulting from individual-level AI provision.

Eleanor W. Dillon
Microsoft Research
eldillon@microsoft.com

Nicole Immorlica
Microsoft Research
nicimm@gmail.com

Sonia Jaffe
Microsoft Research
spj@uchicago.edu

Christopher T. Stanton
Harvard University
Harvard Business School
and NBER
christopher.t.stanton@gmail.com

Shifting Work Patterns with Generative AI*

Eleanor Wiske Dillon[†]

Microsoft Research

Sonia Jaffe[†]

Microsoft Research

Nicole Immorlica

Microsoft Research

Christopher T. Stanton

Harvard Business School,
NBER, and CEPR

November 13, 2025

Abstract

We present evidence from a field experiment across 66 firms and 7,137 knowledge workers. Workers were randomly selected to access a generative AI tool integrated into applications they already used at work for email, meetings, and writing. In the second half of the 6-month experiment, the 80% of treated workers who used this tool spent two fewer hours on email each week and reduced their time working outside of regular hours. Apart from these individual time savings, we do not detect shifts in the quantity or composition of workers’ tasks resulting from individual-level AI provision.

Generative AI has opened new possibilities for technology to assist with or automate a variety of tasks. Early studies have already shown that generative AI increases worker productivity in targeted tasks (e.g. Brynjolfsson, Li and Raymond, 2025; Peng et al., 2023; Noy and Zhang, 2023; Dell’Acqua et al., 2023), and advances in this technology have the potential to shift the nature of jobs (Eloundou et al., 2024; Tomlinson et al., 2025) and perhaps the macroeconomy (Autor, 2024; Acemoglu, 2025). In lab experiments and early field studies on generative AI, participants often focused on completing a single task with targeted, well-suited tools. In everyday settings, workers must first learn which parts of their job benefit from the use of new tools and then change their work habits. This dynamic helps to explain why the transition from the creation of new technologies to effective, widespread adoption in the workplace often takes time (Brynjolfsson, Rock and Syverson, 2021).

In this paper, we present field evidence on the adoption of generative AI tools and their effects on the way knowledge workers do their jobs. Between September 2023 and October 2024, Microsoft partnered with 66 large firms to run a cross-industry field experiment to measure how access to an integrated generative AI tool, Microsoft 365 Copilot (hereafter, Copilot), changed work patterns. Each firm recruited at least 100 workers for random assignment to treatment and control groups. Treated workers received access to this new generative AI tool while controls continued working with their current technologies. Participants in the study came from varied occupations, but all did work that used the Microsoft Office tools into which Copilot was integrated (e.g., emailing, video meetings, chat,

*Corresponding author: eldillon@microsoft.com. [†] denotes equal contribution. We thank the Microsoft Customer Research Program, especially Alexia Cambon, Sida Peng, Neha Shah, Chris Gideon, Modern Work Marketing, the Office of Applied Research, and company partners for help carrying out this experiment, Abigail Atchison, Roman Basko, and Fabio Vera for superb data science support, Jack Cenatempo, Esther Plotnick, and Will Wang for excellent research assistance. We thank 3 anonymous referees, Tatyana Deryugina, Seema Jayachandran, Rem Koning, Anders Humlum, and Danielle Li for helpful comments.

document creation). Firms were asked to maintain the random assignment for six months, giving treated workers time to explore the new tool and integrate it into their daily work.

The study took place during the pilot rollout of Copilot, when firms could purchase licenses for only a small fraction of their workers. Treated workers were free to use or not use the tool during their regular work. Although treated and control workers may have used other generative AI tools (for example, ChatGPT had been available for nearly a year by the beginning of the study), Copilot integrated AI into applications the workers already used regularly in their jobs. This integrated tool allowed new uses, like summarizing email inboxes or searching for local information, and likely reduced the effort to deploy generative AI for writing documents and emails.

We use telemetry data collected by Microsoft products to track workers' behavior during the experiment and for several months beforehand. Our primary outcome measures are related to time use (e.g. time spent in meetings or reading emails) and quantities of activities (e.g., emails read, meetings attended, documents authored) within the Microsoft Office suite, as we do not observe the content of any work nor any measures of productivity or performance evaluation.

Over 90% of the workers assigned to receive Copilot used it at least once in the 6 month period after treatment. Usage peaked within the first 5 weeks of treatment assignment, possibly because of training, novelty, or trial motives. To provide evidence on the effects of AI use when workers were already familiar with the tool, our main analysis focuses on outcomes at least 12 weeks after treatment assignment, where weekly average use stabilized at just under 40% of treated workers.

Workers varied in the intensity of their Copilot use. Firm-level differences explain the largest share of this variation, followed by workers' pre-period behavior. This variation is consistent with other surveys of AI adoption across firms (Bonney et al., 2024; McElheran et al., 2024), but it is notable in our more homogenous sample of large firms that were interested in being early adopters of this tool and participating in a study of its use. Firm characteristics like industry capture only some of the firm effects, suggesting an important role for firm-specific managerial practices or training that we cannot measure (McElheran et al., 2025).

Workers who used Copilot exhibited some meaningful changes in work behavior. Intention-to-treat estimates suggest workers with Copilot reduced their time using Outlook for email by 1.4 hours per week, a 12% reduction relative to the pre-period mean. The Local Average Treatment Effect (LATE) estimate from instrumenting Copilot use with treatment assignment implies a 2 hour (17%) reduction in Outlook time. Copilot users also consolidated when they worked on email, opening up 1.5 hours each week of extended work time without Outlook activity.¹ In contrast, we find no significant average changes in total time spent in video meetings or writing documents.

We see clear evidence that workers did some of their existing work tasks more quickly, but do not detect meaningful shifts in workers' tasks (Autor, 2024) or innovations that reorganize work (Brynjolfsson and Hitt, 2000). In particular, treated and control workers replied to the same number of email threads, participated in the same number of Teams meetings, and completed the same number of Word documents.² While we cannot observe a complete accounting of activities during work, given the occupations of the studied workers and their employment at Office-heavy firms, it would be surprising if they took on major new work responsibilities while leaving no trace of changing behavior in these digital applications. Treated workers did reduce their time spent in digital applications outside of their regular working hours by about 0.25 hours-per-week (9%), suggesting that some of the time saved on email returned to workers. We do not find any negative spillover effects on the work patterns

¹Appendix B.1 details how we define each of these outcomes.

²While not reported, we also find little change in activity in PowerPoint or Excel.

of treated workers’ close coworkers, suggesting that the time savings we document do not come at the expense of work quality.

This study highlights an under-appreciated role for generative AI in real world production processes. In our sample, workers were most likely to access Copilot through Teams or Outlook. The Teams use in particular demonstrates the value of generative AI tools for summarizing and organizing dispersed information. While this role of AI has received less attention than text or code generation, its importance is supported by other recent studies. For example, Tomlinson et al. (2025) identify information gathering as the most common use of a generative AI chat tool. Blandin, Bick and Deming (2024) find that managers, who spend more time in meetings than other workers (Impink, Prat and Sadun, 2025), are tied with computer/mathematical occupations as the groups most likely to use generative AI at work.

Our work adds to studies on the effects of generative AI in real workplaces (Brynjolfsson, Li and Raymond, 2025; Cui et al., 2024; Otis et al., 2023), while complementing studies that use controlled tasks in a lab setting.³ We also contribute to the broader literature on technology diffusion (Comin and Hobijn, 2010; Forman, Goldfarb and Greenstein, 2018). Surveys indicate rapid adoption of generative AI at work across a broader set of workers than those considered in lab or early field settings. Humlum and Vestergaard (2025a) find 41% of Danish workers in exposed occupations used ChatGPT at work by the end of 2023 and Blandin, Bick and Deming (2024) find 24% of all U.S. workers used generative AI at work by August 2024. In follow-up work, Humlum and Vestergaard (2025b) find that AI chatbots at work have not had measurable impacts on worker earnings or hours worked, largely consistent with our mixed results on worker behavior. This study bridges some of the gaps between these two lines of research: measuring both take up and impacts of use for a generative AI tool on a broad set of workers, albeit at the cost of clear measures of worker productivity.

1 Treatment Assignment and Patterns of Copilot Adoption

1.1 Experimental Design

In fall 2023, Microsoft offered select large firms the option to purchase exactly 300 Copilot licenses. Copilot was not available to smaller firms or to individual consumers, and firms could not purchase additional licenses beyond 300. Within this group, Microsoft’s marketing team recruited firms to join the study with the understanding that they would allocate at least 50 of these 300 licenses to the experiment and recruit at least twice as many participating workers. The Copilot licenses were assigned at random among experiment-eligible workers, but treated workers were not required to use the tools or deploy them for specific work tasks. In exchange for participating in the study, firms received early access to findings from the experiment, including customized firm-level reports. They did not receive discounts or access to additional licenses.

Appendix A includes more details on the 7,137 participating workers (3,684 treated) across 66 firms. The firms span a range of industries (see Figure A.1a). All are very large and most operate in multiple countries. Firms opted into the study in response to only modest incentives and may be more proactive in adopting new technologies than the average firm.

Firms varied in how they recruited participating workers, with some targeting specific functions or workers with a track record of adopting new technologies early and others creating an open call

³See, for example, Peng et al. (2023), Spatharioti et al. (2023), Vaithilingam, Zhang and Glassman (2022), Campero et al. (2022), Noy and Zhang (2023), Dell’Acqua et al. (2023), Cambon et al. (2023).

for participation.⁴ Appendix Table B.1 tabulates occupations for the subset of workers for whom we have that information. As expected, the sample is dominated by knowledge workers who regularly do the kinds of emailing and word processing tasks for which Copilot was designed. Appendix Table B.2 summarizes pre-period work behaviors for treated and control workers in the sample. Unfortunately, our data use agreements do not allow us to compare these sampled workers with the broader set of workers at the firms, but we can compare our sample with the knowledge workers studied in Impink, Prat and Sadun (2025). They document workers’ email and meeting behavior in a sample of 102 large firms around 2019. Our sample of workers seems roughly comparable to theirs. For example, their studied workers attended a mean of 39 meetings per month while ours attended 43. Nonetheless, we caveat that participating workers may not be representative of the overall worker population at participating firms, or of knowledge workers more broadly.

1.2 Adoption

The median treated worker used Copilot in 39% of the weeks they had access to it during the study (33% of the weeks in months 4-6).⁵ Figure 1a shows the distribution of individual usage rates over months 4-6; 5% of treated workers used Copilot every week, while 20% never used it at all. As shown in Figure 1c, overall usage peaked at 55% of treated workers in the first weeks of the study as workers explored the new tool and, at some firms, participated in introductory training sessions.

Figure 1c also shows Copilot use separately for Outlook, Teams, and Word. We focus on time use and behavior in these applications because the early phase of prompt-based generative AI was most applicable to tasks in these tools. Workers were most likely to use Copilot in Teams, which can analyze transcripts of meetings to generate summaries, list follow-up items, or answer user queries. The next most common use was through Outlook, where Copilot can summarize email chains, flag emails that require a response, answer questions more flexibly than a keyword search, and generate draft emails in response to prompts. The Word Copilot can draft text from prompts and answer questions about document content. The Word Copilot had the lowest usage of the three we study, partially because not all experimental workers used Word regularly in their jobs (see Appendix B). Our regressions for Word-based outcomes limit the sample to the 2,525 study participants who used Word regularly in the pre-period (39% of our primary analysis sample). Copilots for PowerPoint and Excel had lower usage rates at the time of the study. The application-specific Copilot usage rates displayed in the figure sum to more than the overall usage rate because, as illustrated in Appendix Table C.1, most workers who used Copilot used it in multiple applications. For example, 80% of workers who used Copilot in Outlook in months 4-6 also used Copilot in Teams at least once in the same period.

Firms are a key driver of workers’ propensity to use Copilot. As shown in Figure 1d, usage rates varied widely and consistently across firms, with average usage rates among treated workers as low as 6.3% per week at one firm over the final 3 months of the study and as high as 70% at another. Some of this variation stems from differences across industries. In telecommunications and technology firms, 45% of treated workers used Copilot each week, but only 25% did at construction and manufacturing firms (see Figure 1b). However, considerable variation in usage rates exists within industries.

To better understand the importance of the firm in driving adoption, we test how well we can predict

⁴We suspect that our sample excludes the highest-priority Copilot users, who would have received licenses with certainty rather than by lottery. However, our estimates may still overstate population average treatment effects if firms targeted the experiment to workers where they expected to see large gains.

⁵In our data, use means taking an active action to engage with a Copilot tool through any of the Copilot-enabled applications: Outlook, Teams, Word, Excel, PowerPoint, and the separate M365 Chat. Having Copilot open by default is not sufficient. We do not observe any information about prompts or the type of use, only which application-specific Copilot tools were engaged.

workers’ use of Copilot by using proxies for the type of work done by individual study participants (pre-experiment work patterns), the type of work done at different firms (firm industry and average work patterns), and how firms distributed their Copilot licenses (the share of a worker’s coworkers who had a license). As shown in the first column of Table 1, individual pre-experiment activity in Outlook, Teams, and Word explains only 9.2% of the variation in Copilot adoption for all treated workers in months 4-6. Pre-experiment emails read have a small negative association with the share of weeks an individual used Copilot (the coefficient is per thousand emails), whereas time spent and activity in both Teams and Word are positively associated with use. Email and Teams meetings are positively correlated, making coefficient interpretation challenging, but the implied usage effect of a standard deviation change in Teams meetings (see Table 2 for summary statistics) tends to be larger than the implied effect of a standard deviation change in emails read or threads replied to. The share of coworkers that have Copilot, capturing firms’ strategy in deploying these early licenses, adds little explanatory power (Column 2).⁶ Columns 3 and 4 include firm-level averages of the same pre-period work behavior (coefficients are reported in Appendix Table C.2). They explain 8.7% of the variation in adoption alone, with explanatory power loading on similar behaviors, and 12% with the individual-level averages.

Firm fixed effects alone explain nearly twice as much variation in individual workers’ Copilot use as workers’ pre-period work behavior (comparing Table 1 Columns 5 and 1). Combining the two sets of predictors explains 21% of the variation in adoption rates (Column 6). As suggested by Figure 1, industry fixed effects (Column 7) explain only 2.9% of the variation, far less than firm fixed effects. Industry fixed effects and indicators for the month each firm entered the study explains 8% of the variation (Column 8). However, as shown in Appendix Table C.2, these start month indicators do not follow any predictable pattern.⁷ The final column of Table 1 includes individual workers’ behavior and all observable firm characteristics except the indicators for firm identity. Together, these factors explain 15% of the variation in Copilot use, still less than firm fixed effects alone. Overall, these patterns suggest an important role for differences across firms, such as managerial practices or organizational philosophy, that have been highlighted in earlier studies but are unobservable for us.

2 Direct Effects of Generative AI on Patterns of Work

2.1 Methods

Our main estimation is a difference-in-differences (DiD) approach, which uses pre-experiment data to account for baseline individual heterogeneity and increases precision relative to a simple comparison of means. Using worker i by week t level data, our intent-to-treat (OLS) estimates come from:

$$Y_{it} = \alpha_i + \delta_{m_t} + \gamma_{f,\tau_{ft}} + \beta^{ITT} Z_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \epsilon_{it} \quad (1)$$

where f is the firm, α_i is an individual fixed effect, δ_{m_t} captures calendar time (year-by-month) effects, $\gamma_{f,\tau_{ft}}$ are firm-by-event month fixed effects, Z_i is the worker’s experimental treatment assignment, and τ_{ft} is the event month for firm f . The estimation sample includes only workers in the experiment. Firm-event month fixed effects absorb firm-specific trends and fluctuations in work patterns. Calendar

⁶The .013 coefficient implies that moving from no coworkers with Copilot to the median of 11.9% increases the predicted share of Copilot usage weeks by 0.15 percentage points.

⁷If, for example, firms who entered the study later had access to better versions of the Copilot tools that affected uptake, we would expect a monotone pattern, which we do not find. Instead, industry and starting month combinations appear to serve as noisy proxies for firm fixed effects.

time fixed effects modestly aid precision by further capturing common seasonal trends.⁸ We cluster standard errors at the individual level.

With this design, the effects of Copilot are identified by differences in the month-to-month changes in outcomes between treated workers and comparison workers at the same firm. The potential biases from a rolling DiD (Goodman-Bacon, 2021) are therefore not a concern. Our baseline estimates omit months 0-3 (weeks 0-12) after each firm’s experimental licenses were assigned to focus on the longer-term effects of Copilot use.⁹

We also estimate local average treatment effects by instrumenting a Copilot use indicator with treatment assignment. Instrumenting overcomes issues with non-compliance in both directions: workers who are assigned a Copilot license but do not use the tool and the less common (but non-trivial) case of workers who were assigned to the control group but nonetheless obtain Copilot licenses (Duflo, 2001; Hudson, Hull and Liebersohn, 2017). The outcome equation is

$$Y_{it} = \alpha_i + \delta_{m_t} + \gamma_{f, \tau_{ft}} + \beta^{IV} D_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \epsilon_{it} \quad (2)$$

where $D_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\}$ is one in the post-period for individuals that used Copilot in months 4 through 6 and zero otherwise. We estimate the model via two-stage least squares, instrumenting $D_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\}$ with $Z_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\}$.

Because we test multiple hypotheses, we use sharpened q -values that account for the increased false discovery rate as the number of tests increases (Anderson, 2008). The q -values (reported as asterisks in tables) can be interpreted like p -values after accounting for the total number of hypotheses tested. The number of tests reflect all parameter estimates for Tables 2 and 3 in the main text, along with the additional outcomes considered in Appendix Table D.2.

2.2 Results

We focus on two broad sets of outcomes: 1) time use in digital applications, and 2) quantity-related metrics such as the number of meetings attended and documents completed. The time use measures for Outlook and Word are constructed from timestamps associated with user actions and defined as sessions that capture closely-spaced actions. An Outlook session begins when a user opens an email in the reading pane, continues as long as the user records another “reading” action at least every 15 minutes, and ends when the user exits the reading pane and does not open another email for 15 minutes or more. Word sessions are similarly based on 15 minute gaps in activities (e.g., interactions with a document or series of documents).¹⁰ The quantity measures are straightforward, as is the time spent attending Teams meetings. Table 2 contains the main results.

2.2.1 Effects on Time Use

The most prominent impact of Copilot on worker behavior that we observe involves changes in how workers manage email. In the pre-period, the average worker spent 11.7 hours per week reading and

⁸Both effects are identified because the event-time indicators mark 4-week blocks following each firm’s start date, which may fall mid-calendar month.

⁹Estimates using all post-treatment weeks are in Appendix Table D.1.

¹⁰Importantly, we do not have complete measures of workers’ actions composing and sending email, so time spent creating emails must be inferred as the time between interactions with the reading pane. The Outlook session metrics were developed in consultation with the Microsoft product team that manages the M365 telemetry data. Patterns of Outlook and Word session time are robust to cutoffs of 10 or 20 minutes rather than 15. Workers may multi-task or switch rapidly between applications such that sessions in different applications overlap. Due to data processing challenges, we do not measure or adjust for this overlap across apps, so session-based measures may overstate the actual time spent on each activity. Details on the construction of all worker behavior measures are in Appendix B.1.

responding to emails. Our intent-to-treat (LATE) estimates show workers with access to Copilot spent an average of 1.4 (2) fewer hours per week in email sessions over post-period months 4-6, a 12% (17%) decline from their pre-period average. Workers with access to Copilot also consolidated their email work, reducing their count of Outlook sessions per week by 11% (17%), while opening up almost 1.5 hours per week of time in “non-email” blocks of 30 minutes or more during the regular workday with no Outlook activity (LATE = 2.2 hours). These email-free blocks capture longer periods where workers may be able to focus on other projects without interruptions from email (Jackson, Dawson and Wilson, 2001).

We find no similar shifts in the time workers spent in Teams meetings or working in Word, despite the fact that workers regularly used Copilot through these applications. The ITT results rule out effect sizes outside of the interval between -0.01 and 0.21 hours for Teams meetings relative to a mean of 5.22 hours per week. We estimate an increase in Word session time of 0.14 (LATE = 0.19) hours per week, an 8.6% increase from a small pre-period mean, but this effect just misses statistical significance using the sharpened q -values. The smaller sample of regular Word users leaves us somewhat under-powered to detect changes in document-related work.¹¹

Across applications, the last row of the table shows workers spent a quarter of an hour less per week working outside of their modal work hours, a 9% reduction, with somewhat larger time savings in after-hours email work. Together, these results suggest that workers used some of their time saved on email to shorten their work day, but much of the time saved appears to have shifted to activities we cannot measure. The dynamic difference-in-differences results in Figure 2 illustrate that the effects on Outlook time are relatively stable over the full post-period, with persistent effects that are apparent even in the first month after workers had access to the tool. The Teams and Word results remain close to zero in all months.

Although our treatment comprises access to, or use of, Copilot through any Office application, the Outlook time savings likely come from Copilot use in Outlook rather than cross-application effects. While transcription tools in Teams Copilot could, in principle, reduce the need to write follow-up emails, reduced form regressions of outcomes on application-specific Copilot use (see Appendix Table D.3) do not find much correlation between Teams Copilot use and changes in outcomes for email or document completion after controlling for use of other Copilots.¹²

Given the differences in Copilot adoption in Table 1, a natural question is whether treatment effects also vary by group. We use the machine learning procedures in Chernozhukov et al. (2025) to evaluate heterogeneity in conditional average treatment effects (CATEs) across a high-dimensional set of covariates. We cannot reject the null of 0 treatment effect heterogeneity for total Outlook session time when we allow the model to choose which dimensions of potential heterogeneity matter via elastic net or a neural networks approach (see Appendix Table D.4 for details and estimates). Group average treatment effects for the top and bottom quintiles of workers by CATE are only modestly different than the full average treatment effects in Table 2 (-1.65 and -1.18 hours for the most and least affected workers) and we cannot reject that treatment effects for these two groups are equal.

¹¹We believe differences in Word use reflect the range of jobs done by our sample of workers. See Word usage by occupation in Appendix B.2.

¹²Note that Copilot use is strongly correlated across applications (see Appendix Table C.1), limiting our power to detect cross-application relationships.

2.2.2 Effects on Task Quantity and Type

In addition to consolidating their email work into fewer sessions per week, we find workers with Copilot also generated time savings by reading 4% (LATE = 6%) fewer emails each week (Table 2)¹³ with no increase in their time spent per email (in Appendix Table D.2 we rule out increases larger than 1.2%). Table 2 shows no significant changes in the number of unique threads replied to or in the time from delivery to reply (we can rule out effect sizes greater than 3.4% of the mean in the ITT estimates for email threads and greater than 4.1% of the mean for reductions in time to reply to emails), suggesting that workers were not ignoring their inboxes or missing emails that required a reply. The number of emails per thread (Appendix Table D.2) did not change, suggesting that Copilot may have been used for email summaries but with limited effects on the mix of emails or on thread efficiency. Given the estimated time savings, workers’ responses appear to reflect their ability to take less time to do roughly the same quantity of email work.

The null effects of Copilot on time spent in Teams meetings or Word document writing could reflect offsetting responses to a productivity-enhancing tool: workers complete the same work in less time but use the time saved to do more of that kind of work. These competing shifts are akin to the income and substitution effects generated by a change in price — in this case, the amount of time required to complete a task. If there were offsetting or counterbalancing effects, we would expect to see quantity responses, such as an increase in the number of meetings or documents. We might also expect shifts in the composition of tasks, like observing more short meetings or fewer recurring meetings. However, we find no evidence for these kinds of changes for Teams tasks in Table 2 (we can rule out increases in the number of meetings greater than 4.6%).¹⁴

The patterns in the third panel of Table 2 are consistent with these two counterbalancing effects, but the estimates are imprecise and the point estimates are small. Regular Word users with Copilot completed their documents more quickly, consistent with Noy and Zhang (2023), but the ITT point estimate is a savings of only 2% of the pre-period mean, and we cannot rule out that document time to complete increased by as much as 7%.¹⁵ We estimate that Copilot use also slightly increased the total Word documents authored by the studied workers (we can rule out increases larger than 11% of the pre-period mean and cannot rule out no change). Time to complete fell the most for collaborative documents (over 2 days, or a 17% reduction), however even these effects are not statistically significant using the sharpened q -values.¹⁶ Any time savings are not accompanied by an increase in collaborative document production or the share of authored documents that include collaborators.

3 Work Quality, Spillovers, Group Effects, and Co-Invention

The patterns in Table 2 suggest that workers with Copilot took less time to do roughly the same quantity of email work. Without observing the content of the emails they write, we cannot say definitively whether this time savings came at the cost of reduced work quality. However, we can

¹³Our measure counts unique emails read each week. If a worker revisited an email, those actions would count as email reads in each week. If a worker only used Copilot to summarize information, that would not count as an email read.

¹⁴As shown in Appendix Table D.2, we find one statistically significant compositional change: a small (0.06 to 0.085 per week) increase in the count of short meetings with many participants. Rogelberg (2019) finds that large meetings are most effective for information transmission rather than ideation or problem solving. Copilot may have swayed busy workers who previously skipped some of these meetings to attend instead because they could transcribe and catch up later, even if multitasking in the moment.

¹⁵Note that the time to complete a document is distinct from the total time spent in Word. Time to complete is based on the elapsed time from when a document is created to its approximate end (when 90% of all edits are done), whereas Word session time measures the actual time users spend in Word.

¹⁶A document is defined as collaborative if a worker was the primary editor but at least one other worker contributed. Non-collaborative documents are individual.

provide suggestive evidence by considering spillover effects on the colleagues of treated workers. As discussed in Appendix B.3, we identify close colleagues as the coworkers one interacts with most frequently through emails, meetings, and joint writing. If treated workers were inefficiently reducing time spent on email or sending lower quality replies generated by Copilot, we might expect to see their colleagues spending more time in email to pick up the slack. Alternatively, treated workers may have established effective new practices that also benefited their teams.

To test these channels, we estimate a reduced-form intent-to-treat model that captures potential spillovers. In doing so, we are also able to account for potential violations of the Stable Unit Treatment Value Assumption (SUTVA) because of spillovers to the control group and to consider whether Copilot provision to a close coworker changes treatment effects for the focal worker. The regression is

$$Y_{it} = \alpha_i + \delta_{m_t} + \gamma_{f, \tau_{ft}} + \beta^{ITT} Z_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \beta^{Close} Z_{Close} \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \beta^{ITT \cdot Close} Z_i \cdot Z_{Close} \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \epsilon_{it} \quad (3)$$

where the model mimics equation (1), adding the indicator for having a close connection who is offered Copilot through the experiment, Z_{Close} , and an interaction of the focal worker’s treatment and Z_{Close} . As shown in Appendix Figure B.1, 30% of studied workers have at least one treated coworker during the post-period.¹⁷

The first column of Table 3 shows the main treatment effect estimates when we control for potential spillovers to the control group through Z_{Close} , a strategy that De Grip and Sauermann (2012) and Espinosa and Stanton (2022) use to assess the impact of potential SUTVA violations. The point estimates are very similar to the reduced form estimates in Table 1, have the same patterns of statistical significance based on the sharpened q-values, and generally fall within the confidence intervals of the original estimates.

It should not be surprising, given the stability of estimates for the focal worker, that the results in column two for having a close co-worker treated are generally much smaller in magnitude than the main effects. The spillover effects on coworkers’ email activity generally have the same sign as the main effects. However, adjusting for multiple hypothesis testing, the only statistically significant estimate is on unique email threads replied to, which falls. Overall, these spillover estimates suggest that treated workers were using Copilot to reduce email time in efficient ways, with modest benefits for their coworkers. This provides suggestive evidence that their gains did not come at the expense of quality.

In keeping with the largely null direct effects of Copilot for meetings or writing tasks, there are no spillover effects of Copilot access on connected coworkers’ Teams or Word activity. These consistent differences across applications may reflect how easily individual workers can adjust their work to take advantage of the new tool. Email is solitary and each worker is free to develop her own process for inbox management. In contrast, shifting what work gets done in meetings, how long they last, or who takes primary responsibility for writing documents requires coordinating with colleagues and agreeing on new norms. This pattern is fitting for the early adoption stage we study, where individual workers were exploring the capacities of a new tool, but the complementary process innovations and organizational restructuring that drive larger changes in workflow had not yet taken place (Bresnahan et al., 1996; Brynjolfsson and Hitt, 2000).

¹⁷Workers may also have close colleagues who received one of the firms’ discretionary (i.e. not randomly assigned through the experiment) Copilot licenses. Under our data use agreements, we are unable to track this Copilot access for workers outside the study.

In the spirit of the literature on complementary innovation and process restructuring, we test whether there are different effects when multiple coworkers get access to the tool. The third column of Table 3 shows the triply interacted effects of Z_i , Z_{Close} , and $\mathbf{1}\{\tau_{ft} \geq 0\}$. The total effect of being treated oneself while having at least one treated close colleague is given by the sum across the three columns. The point estimates suggest some modest benefits to gaining access to the new tool along with colleagues, treated workers with treated colleagues save 50% more time on email, but again these effects are imprecisely estimated and, with the sharpened q values, we cannot rule out zero differences in effect.¹⁸ These results suggest that the co-creation of work and process restructuring that the literature identifies as a pre-condition for effective technology adoption did not seem to take place under the individual-level rollout of the tool that our experiment assesses.

4 Conclusion

This paper offers a portrait of the first steps of adoption of a new technology. Historically, the main productivity gains from innovative technologies came only with the complementary process innovations that shift the nature of work around these new capacities. These institutional changes take time. Among recent technological innovations, generative AI is unique because it does not require new equipment or hardware to use, allowing individual workers to bring generative AI to work even without their employers’ support. As Blandin, Bick and Deming (2024) point out, the discrepancy between the high rates of generative AI use reported in surveys of workers, such as theirs, and the lower rates reported in surveys of firms, as in Bonney et al. (2024), suggests that workers use these tools even before their employers develop formal adoption strategies.

This study provides indications of how this individual use of generative AI shapes the patterns of work for knowledge workers in large firms. Large firms are often among the earliest adopters of digital technologies, presumably because of their ability to capture benefits (McElheran et al., 2024). We study workers during the first year in which generative AI tools gained broad use. While their colleagues may have been using other generative AI tools, the workers we studied were often the only members of their team with access to Copilot during the early roll-out period. Although we cannot measure productivity, the changing behaviors we see are consistent with workers independently exploring these new tools and saving time on individual tasks. Further research is needed to assess how co-inventions or team- or firm-level transformations may lead to broader changes, such as shifts in what tasks workers do.

¹⁸The specification here is a 0-1 indicator for whether any close coworker is treated, while a saturation design would account for the share of workers on a team that are treated. Using that approach, reported in Appendix Table D.5, we also detect little difference in treatment effects by whether our studied workers had an above or below median share (11.9%) of close coworkers with access to Copilot. However, even our “high saturation” teams had, on average, only 39% of coworkers with access to Copilot, so we may simply be unable to detect team-level effects from this experimental design.

References

- Acemoglu, Daron.** 2025. “The simple macroeconomics of AI.” *Economic Policy*, 40(121): 13–58.
- Anderson, Michael L.** 2008. “Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” *Journal of the American Statistical Association*, 103(484): 1481–1495.
- Autor, David.** 2024. “Applying AI to rebuild middle class jobs.” National Bureau of Economic Research.
- Blandin, Adam, Alexander Bick, and David Deming.** 2024. “The Rapid Adoption of Generative AI.” Available at SSRN 4965142.
- Bonney, Kathryn, Cory Breaux, Cathy Buffington, Emin Dinlersoz, Lucia S Foster, Nathan Goldschlag, John C Haltiwanger, Zachary Kroff, and Keith Savage.** 2024. “Tracking firm use of AI in real time: A snapshot from the Business Trends and Outlook Survey.” National Bureau of Economic Research.
- Bresnahan, Timothy, Shane Greenstein, David Brownstone, and Ken Flamm.** 1996. “Technical progress and co-invention in computing and in the uses of computers.” *Brookings Papers on Economic Activity. Microeconomics*, 1996: 1–83.
- Brynjolfsson, Erik, and Lorin M Hitt.** 2000. “Beyond computation: Information technology, organizational transformation and business performance.” *Journal of Economic perspectives*, 14(4): 23–48.
- Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond.** 2025. “Generative AI at work.” *The Quarterly Journal of Economics*.
- Brynjolfsson, Erik, Daniel Rock, and Chad Syverson.** 2021. “The productivity J-curve: How intangibles complement general purpose technologies.” *American Economic Journal: Macroeconomics*, 13(1): 333–372.
- Cambon, Alexia, Brent Hecht, Benjamin Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan.** 2023. “Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity.” Microsoft.
- Campero, Andres, Michelle Vaccaro, Jaeyoon Song, Haoran Wen, Abdullah Almaatouq, and Thomas W. Malone.** 2022. “A Test for Evaluating Performance in Human-Computer Systems.”
- Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val.** 2025. “Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India.” *Econometrica*, 93(4): 1121–1164.
- Comin, Diego, and Bart Hobijn.** 2010. “An exploration of technology diffusion.” *American economic review*, 100(5): 2031–2059.

- Cui, Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz.** 2024. “The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers.” *Available at SSRN 4945566*.
- De Grip, Andries, and Jan Sauermann.** 2012. “The effects of training on own and co-worker productivity: Evidence from a field experiment.” *The Economic Journal*, 122(560): 376–399.
- Dell’Acqua, Fabrizio, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani.** 2023. “Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality.” *Harvard Business School Technology & Operations Mgt. Unit Working Paper*.
- Duflo, Esther.** 2001. “Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment.” *American economic review*, 91(4): 795–813.
- Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock.** 2024. “GPTs are GPTs: Labor market impact potential of LLMs.” *Science*, 384(6702): 1306–1308.
- Espinosa, Miguel, and Christopher T Stanton.** 2022. “Training, communications patterns, and spillovers inside organizations.” National Bureau of Economic Research.
- Forman, Chris, Avi Goldfarb, and Shane Greenstein.** 2018. “Technology adoption.” In *The Palgrave Encyclopedia of Strategic Management*. 1719–1723. Springer.
- Goodman-Bacon, Andrew.** 2021. “Difference-in-differences with variation in treatment timing.” *Journal of econometrics*, 225(2): 254–277.
- Hudson, Sally, Peter Hull, and Jack Liebersohn.** 2017. “Interpreting instrumented difference-in-differences.” *Metrics Note, Sept.*
- Humlum, Anders, and Emilie Vestergaard.** 2025a. “The unequal adoption of ChatGPT exacerbates existing inequalities among workers.” *Proceedings of the National Academy of Sciences*, 122(1): e2414972121.
- Humlum, Anders, and Emilie Vestergaard.** 2025b. “Large Language Models, Small Labor Market Effects.” *University of Chicago, Becker Friedman Institute for Economics Working Paper*.
- Impink, Stephen Michael, Andrea Prat, and Raffaella Sadun.** 2025. “Communication within firms: evidence from CEO turnovers.” *Management Science*, 71(1): 470–487.
- Jackson, Thomas, Ray Dawson, and Darren Wilson.** 2001. “The cost of email interruption.” *Journal of systems and information technology*, 5(1): 81–92.
- McElheran, Kristina, J Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas.** 2024. “AI adoption in America: Who, what, and where.” *Journal of Economics & Management Strategy*, 33(2): 375–415.
- McElheran, Kristina, Mu-Jeung Yang, Zachary Kroff, and Erik Brynjolfsson.** 2025. “The Rise of Industrial AI in America: Microfoundations of the Productivity J-curve (s).”
- Noy, Shakked, and Whitney Zhang.** 2023. “Experimental evidence on the productivity effects of generative artificial intelligence.” *Science*, 381(6654): 187–192.

- Otis, Nicholas, Rowan P Clarke, Solene Delecourt, David Holtz, and Rembrand Koning.** 2023. “The uneven impact of generative AI on entrepreneurial performance.” *Available at SSRN 4671369*.
- Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer.** 2023. “The impact of AI on developer productivity: Evidence from GitHub copilot.” *arXiv preprint:2302.06590*.
- Rogelberg, Steven G.** 2019. *The surprising science of meetings: How you can lead your team to peak performance*. Oxford University Press.
- Spatharioti, Sofia Eleni, David M Rothschild, Daniel G Goldstein, and Jake M Hofman.** 2023. “Comparing traditional and LLM-based search for consumer choice: A randomized experiment.” *arXiv preprint:2307.03744*.
- Tomlinson, Kiran, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri.** 2025. “Working with AI: Measuring the Occupational Implications of Generative AI.” *arXiv preprint arXiv:2507.07935*.
- Vaithilingam, Priyan, Tianyi Zhang, and Elena L Glassman.** 2022. “Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models.” *Chi conference on human factors in computing systems extended abstracts*, 1–7.

Tables and Figures

Table 1: Predicting Adoption

Dependent Variable:	Individual average Copilot use								
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.282 (0.011)	0.279 (0.012)	0.142 (0.020)	0.116 (0.022)					
Emails read (thousands)	-0.207 (0.060)	-0.203 (0.060)		-0.183 (0.066)		-0.198 (0.063)			-0.191 (0.065)
Unique email threads replied to (thousands)	-0.853 (0.559)	-0.862 (0.559)		-0.572 (0.588)		-0.572 (0.567)			-0.492 (0.578)
Teams meeting time (hours)	0.012 (0.003)	0.012 (0.003)		0.011 (0.003)		0.011 (0.003)			0.011 (0.003)
Total Teams meetings attended	0.007 (0.002)	0.007 (0.002)		0.002 (0.002)		0.002 (0.002)			0.002 (0.002)
Word total session time (hours)	0.017 (0.005)	0.017 (0.005)		0.010 (0.005)		0.011 (0.005)			0.011 (0.005)
Others' documents read/edited before completion	0.038 (0.013)	0.037 (0.013)		0.031 (0.013)		0.030 (0.013)			0.031 (0.013)
Number of email recipients (tens)	-0.002 (0.004)	-0.002 (0.004)		-0.006 (0.004)		-0.005 (0.004)			-0.006 (0.004)
Number of unique people met 1-1 (total, tens)	0.003 (0.005)	0.003 (0.005)		0.003 (0.005)		0.003 (0.005)			0.002 (0.005)
Share of coworkers with Copilot license		0.013 (0.023)		0.054 (0.023)		0.051 (0.024)			0.051 (0.023)
Firm pre-period behavior			Yes	Yes					Yes
<i>Fixed-effects</i>									
Firm					Yes	Yes			
Industry							Yes	Yes	Yes
Experiment start calendar month								Yes	Yes
Observations	3,422	3,422	3,684	3,422	3,684	3,422	3,684	3,684	3,422
R ²	0.092	0.092	0.087	0.120	0.181	0.213	0.029	0.081	0.151

Note: This table reports OLS regressions where the dependent variable is the share of weeks in post-period months 4-6 that an individual treated worker used Copilot. Each treated worker contributes a single observation. Because 80% of treated individuals used Copilot at least once in this period, most of the variation is on the intensive margin. Firm pre-period behavior controls (Columns 3 and 4) comprise the means of the individual-level regressors aggregated to the firm level. See Appendix B.1 for variable definitions and Appendix Table C.2 for estimated coefficients on firm behaviors and start month indicators. Standard errors are heteroskedasticity-robust.

Table 2: Effects of Copilot Provision and Takeup

Metric	(1) Pre-period mean (SD)	(2) OLS (SE)	(3) IV (SE)	(4) Workers	(5) n
Outlook total session time	11.65 (6.81)	-1.37** (0.13)	-2.03** (0.19)	6441	250145
Outlook sessions	31.67 (11.08)	-3.56** (0.30)	-5.30** (0.44)	6441	250145
Email-free work time	26.18 (6.77)	1.48** (0.14)	2.20** (0.21)	6441	250145
Outlook out-of-hours session time	2.19 (1.84)	-0.33** (0.038)	-0.49** (0.057)	6441	250145
Emails read	160.83 (117.74)	-6.69** (1.62)	-10.02** (2.43)	6387	223702
Unique email threads replied to	14.51 (14.17)	-0.16 (0.17)	-0.24 (0.25)	6441	267887
Time to reply from email delivery	16.55 (5.34)	-0.33 (0.17)	-0.47 (0.24)	6432	235121

Teams meeting time	5.22 (3.65)	0.10 (0.054)	0.14 (0.076)	6170	155003
Total Teams meetings attended	9.93 (6.93)	0.24 (0.11)	0.34 (0.15)	6170	253723
Share of mtgs. attended early/late (%)	30.65 (11.94)	0.019 (0.27)	0.026 (0.37)	6086	128660
Recurring Teams meeting time	2.32 (1.93)	0.037 (0.029)	0.052 (0.041)	6170	202314

Word total session time	1.63 (1.58)	0.14 (0.065)	0.19 (0.089)	2525	104913
Avg. document time to complete	186.54 (94.20)	-3.98 (8.96)	-5.27 (11.86)	2521	29683
Avg. collaborative document TTC	287.85 (127.78)	-48.03 (21.50)	-60.56 (27.17)	1763	8700
Avg. non-collaborative document TTC	158.07 (89.13)	0.16 (9.07)	0.22 (12.02)	2506	25252
Completed documents	0.76 (0.80)	0.029 (0.025)	0.039 (0.035)	2525	77138
Completed collaborative documents	0.14 (0.21)	0.011 (0.0072)	0.015 (0.0098)	2525	77138

Total out-of-hours session time	2.72 (2.16)	-0.25** (0.050)	-0.36** (0.075)	6304	244944

Significance: ** $q < 0.01$, * $q < 0.05$

Note: For each outcome variable, this table shows pre-period mean and cross-worker standard deviations (Column 1), the intent-to-treat coefficients for the effect of Copilot (Column 2), and the LATE coefficients for Copilot adopters (Column 3). All time metrics are in hours. There are some differences in worker counts across metrics due to differences in data collection and storage policies across sources, averages not being meaningful when counts are zero, and our exclusion of workers with low/zero pre-period activity for Word and Teams. See Appendix B for details. Share of meetings attended late/early refers to the share of Teams meetings that a worker joined more than 5 minutes after the meeting started or left more than 5 minutes before it ended. We assign documents to a worker if they were the “primary editor,” making the most edits to the document. Time to complete (TTC) and completed document counts are calculated only for documents where the target worker is the primary editor. Standard errors are clustered at the individual level. Asterisks reference sharpened q -values that adjust for the false discovery rate, as in Anderson (2008).

Table 3: Spillover Effects of Copilot

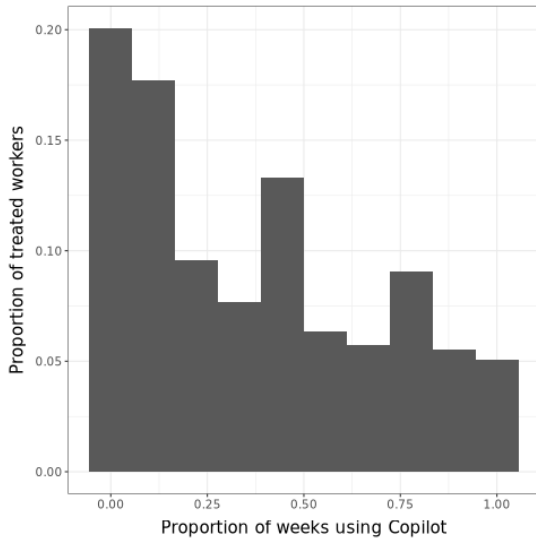
Metric	(1) Direct treatment (SE)	(2) Teammate treatment (SE)	(3) Direct \times Teammate treatment (SE)
Outlook total session time	-1.26** (0.16)	-0.37 (0.17)	-0.29 (0.26)
Outlook sessions	-3.32** (0.36)	-0.54 (0.38)	-0.68 (0.60)
Email-free work time	1.38** (0.17)	0.41 (0.18)	0.28 (0.28)
Outlook out-of-hours session time	-0.28** (0.047)	-0.060 (0.054)	-0.13 (0.077)
Emails read	-7.09** (1.95)	-3.59 (2.34)	1.00 (3.36)
Unique email threads replied to	-0.35 (0.20)	-0.73* (0.26)	0.52 (0.36)
Time to reply from email delivery	-0.16 (0.21)	0.097 (0.26)	-0.47 (0.34)

Teams meeting time	0.17 (0.069)	-0.015 (0.084)	-0.19 (0.11)
Total Teams meetings attended	0.35 (0.14)	-0.35 (0.17)	-0.29 (0.23)
Share of mtgs. attended early/late (%)	0.039 (0.36)	0.38 (0.41)	-0.059 (0.55)
Recurring Teams meeting time	0.077 (0.036)	-0.050 (0.048)	-0.11 (0.061)

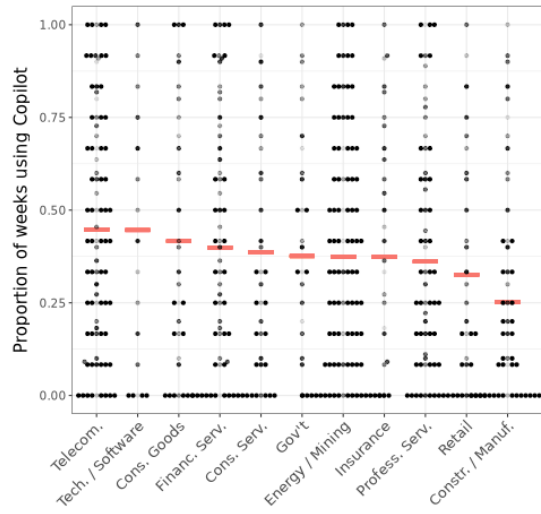
Word total session time	0.084 (0.083)	-0.066 (0.10)	0.16 (0.13)
Avg. document time to complete	-11.86 (11.03)	-11.06 (13.65)	21.07 (18.26)
Avg. collaborative document TTC	-67.20 (29.03)	-26.75 (30.34)	43.06 (43.94)
Avg. non-collaborative document TTC	-0.63 (11.05)	4.81 (13.71)	2.72 (18.53)
Completed documents	0.038 (0.032)	0.053 (0.040)	-0.025 (0.053)
Completed collaborative documents	0.011 (0.0087)	0.0028 (0.011)	-0.00079 (0.015)

Total out-of-hours session time	-0.22** (0.064)	0.034 (0.074)	-0.074 (0.10)

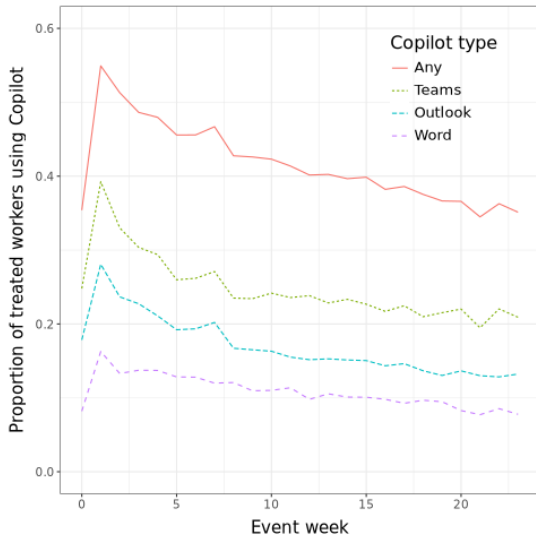
Note: This table shows estimates resulting from a triply-interacted version of the main OLS regression specification, where the three treatments of interest are being in the treatment group, having a close coworker who received Copilot through the experiment, and being in the treatment group while also having a treated coworker. Reported coefficients are the interactions of these terms with the post-period indicator. Asterisks reference sharpened q -values that adjust for the false discovery rate, as in Anderson (2008).



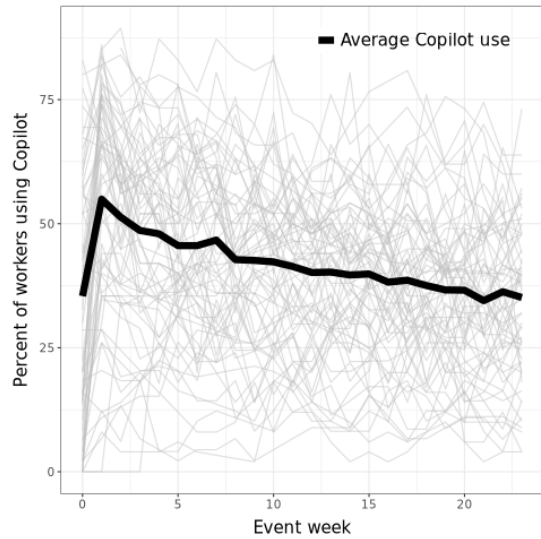
(a) Usage Rate by Individual



(b) Usage Rate by Industry



(c) Usage Rate by App



(d) Usage Rate by Firm

Figure 1: Copilot Usage in the Treatment Group

Note: Panels (a) and (b) show the proportion of weeks in post-period months 4-6 that each worker used Copilot overall (a) and by industry (b). In Panel (b), each solid dot represents 8 workers; lighter dots represent fewer than 8 workers. Mean industry usage rates are highlighted in red. Panel (c) shows the share of treated workers using Copilot in each post-period week, through any application and specifically through Teams, Outlook, and Word. Panel (d) the share of treated workers using Copilot in each post-period week separately for each participating firm. The black line in panel (d) shows the average across firms, replicating the red line in panel (c).

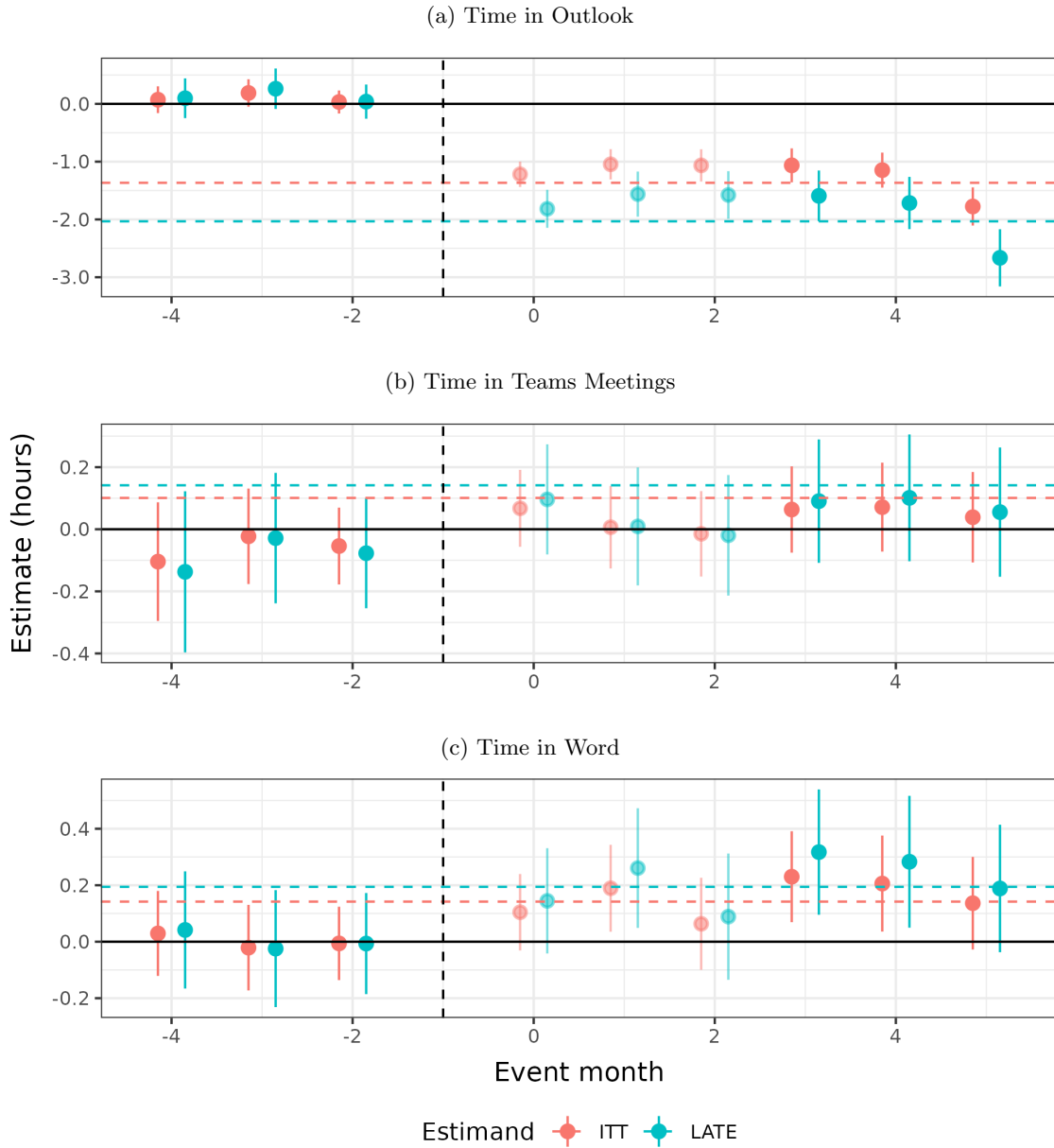


Figure 2: Event Study Plots for the Effect of Copilot on Time Allocation

Note: These plots show coefficients from event-study versions of the regressions defined in Equations (1) and (2). For the ITT estimates, we estimate event month-specific treatment effects, β_τ using $Y_{it} = \alpha_i + \delta_{m_t} + \gamma_{f,\tau_{ft}} + \sum_{k \neq -1} \beta_\tau Z_i \cdot \mathbf{1}\{\tau_{ft} = k\} + \epsilon_{it}$. To extend the IV to a multi-period event study, we interact D_i with event month and instrument with interactions of Z_i and event month. While outcomes are measured on a weekly basis, time-varying effects are at a monthly frequency to reduce noise. Vertical lines show 95% confidence intervals based on individual-clustered errors. Horizontal dashed lines give the average estimated effect from Table 2. These plots show treatment effects by event time, which pools calendar months across firms in our staggered adoption setting. Appendix Table ?? reports treatment effects separately for early- and late-adopting firms.

Supplemental Appendix

A Experiment Details

The experiment was overseen by Microsoft’s marketing group in collaboration with firms. Participating firms chose how they wanted to select workers for the study and were responsible for all communication with participating workers, including recruiting and obtaining consent. Firms shared a list of anonymized identifiers for the participating workers with the experiment organizers at Microsoft. The Microsoft side randomized treatment assignment within this list of workers, shared the assignment back to the firm, and used the anonymized identifiers to track telemetry data for the studied workers. The authors of this paper received permission from Microsoft marketing and the participating firms to use the data collected through the experiment for research.

Participating firms were told that they would need to allocate at least 50 of their 300 Copilot licenses at random and would need to recruit at least twice as many workers to join the experiment in order to populate treatment and control groups. For example, a firm that chose to allocate 150 licenses to the experiment would need to recruit 300 workers. One firm had more than 500 workers participate in the study (that is, nearly all their initial package of licenses were allocated through the study), while another defied the suggested minimum and had only 68 workers participate. Across firms, the average number of treated workers was 55.8, with a standard deviation of 33.1.

The incentive for firms to participate in the experiment was that Microsoft agreed to provide participating firms with early access to the research findings and a firm-level analysis of Copilot usage rates among the study participants and some preliminary effect estimates of Copilot on outcomes such as emails read. Firms received the initial report 2 to 4 months after starting the experiment, usually before the end of the studied period. Some firms may have tried to encourage additional adoption due to low adoption numbers in the report, but we do not see systematic changes around these read-outs in the event studies. Many firms also had access to data on Copilot usage for their own workers through other Microsoft products outside of this experiment.

Our analysis sample includes 66 firms spanning a range of industries tabulated in Figure A.1a. Two additional firms agreed to participate in the study but did not follow through with randomization; these firms are excluded from the analysis sample. The experiment includes workers from many countries, with the majority based in the United States or Europe.

For all firms, our data sample begins on May 1, 2023, providing at least 4 months of data before any Copilot licenses were allocated. We measure the start of treatment for each firm using telemetry data that captures when M365 Copilot licenses were activated for participating workers. When firms alerted Microsoft of the date they allocated licenses, this start date generally aligns with our measured date, but not all firms provided this information. Most firms allocated all experimental Copilot licenses at once, but a few allocated licenses in batches. In a few cases, we observe one or two participating workers receiving Copilot licenses well before the other participating workers at their firm or the stated start date (likely imperfect compliance with the experiment or possibly a data error). For consistency, we set the start date as the day when 25% of the licenses for the experiment were allocated. The earliest start date was in September 2023, and the latest in April 2024. Figure A.1b illustrates the number of firms for whom we are collecting post-activation data over time. We collected data on the workers at each firm for a maximum of 24 weeks after licenses were allocated. 24 weeks is also the modal length of the post period across firms. Five firms have shorter treatment periods because they chose to withdraw from the study before 24 weeks. In all analyses, we drop outcomes for the weeks

ending December 29, 2023 and January 5, 2024 as work activity was low across all workers during this holiday period.

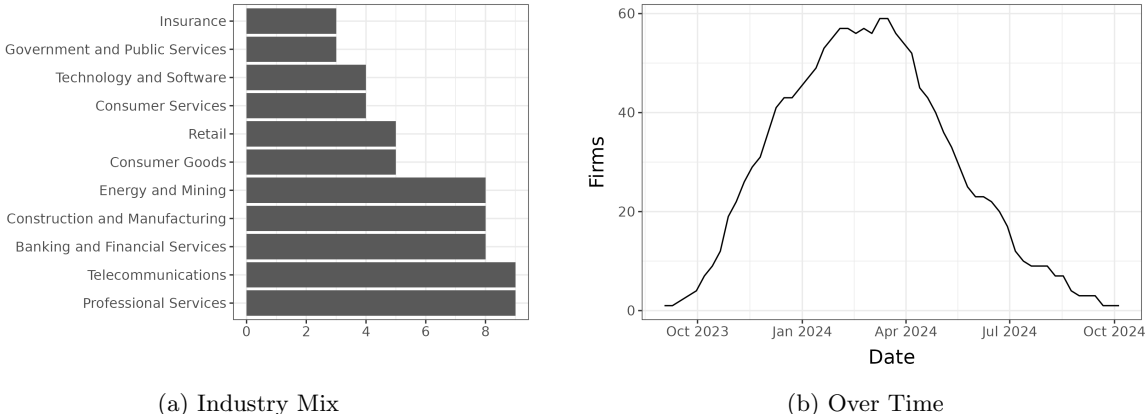


Figure A.1: Participating Firms

Note: Figure (a) shows the number of participating firms in each industry. Figure (b) shows the number of firms active in the experiment over time.

B Data Details

We analyzed hundreds of millions of Microsoft 365 usage signals (telemetry), for both the treated and designated control workers, from product telemetry data that Microsoft collects. To maintain workers’ and firms’ privacy, we do not observe anything about the content produced by these workers, so our focus is on how workers allocate their time and interact with each other rather than on the quality of their work.

One limitation of this study is that we only observe activity within the Microsoft Office products we can track. We are, for example, measuring time spent in meetings through Teams, not time spent in meetings overall. Fortunately, the firms invited to participate in the study are all heavy users of Microsoft products. The vast majority of workers in our sample used Outlook regularly (6,485 out of 7,137) in the pre-period months, and we suspect we are capturing a large share of work-related email activity. Our baseline sample for email outcomes includes 6,441 workers for whom we observed email sessions in the experiment period and any pre-period Outlook use. Our sample differs for some Outlook metrics for reasons we detail in the next subsection.

In contrast, many sample workers show almost no Word activity during our pre-period months. Based on conversations with participating firms and the patterns of application use by occupation in Table B.1, we think this mostly reflects roles that do not involve document creation, though some low activity could reflect use of a different word processing tool. We do not expect (and do not observe) that individual access to Copilot would shift many of these workers towards using Word. If it did, that change from zero to positive activity would have a very different interpretation than the intensive margin behavioral changes of workers with recorded Word activity in the pre-period. We therefore restrict all Word-related analyses to workers with moderate use in the pre-period, indicated by above-median time spent in Word (0.25 hours/week) and above-median number of documents originated as the ‘primary editor’ (0.16 documents/week). Our main Word sample includes 2,525 workers who meet these criteria.

In the case of meetings, workers with near-zero pre-period use of Teams are concentrated in 10 firms, all of whom report primarily using non-Microsoft applications for video meetings. As with document creation, we are unable to establish reliable pre-period baseline meeting activity for these workers. In this case, we believe this largely reflects our inability to measure work meetings rather than true zeros. Any Copilot-induced changes in Teams activity is likely to reflect substitution to Teams rather than true changes in work meetings. We therefore drop workers at these 10 firms for all meeting outcome analyses. Our main Teams sample includes 6,170 workers in the remaining firms.

B.1 Outcome Metrics

We construct weekly metrics capturing each worker’s email, document editing, and meeting behaviors. Metrics are winsorized to the 99th percentile to address telemetry errors and privacy concerns. In the following subsections we define each behavior measured used in the paper. We also note outcome-specific changes in the worker sample or worker-week observation count. Table B.2 shows the balance tests for pre-experiment differences between the treatment and control groups for all outcomes we use in the analysis.

B.1.1 Outlook

For email, we aggregate the signals of various ‘reading actions’ users take into ‘Outlook sessions:’ blocks of time where a user doesn’t go more than 15 minutes without reading an email.¹⁹ Session time begins at the first eligible moment outside a prior session when a user opens an email in the reading pane. It ends either when the user exits the reading pane and doesn’t open an email for at least 15 minutes, with the exact end of the session being the earlier of the user’s exit or 15 minutes after the reading action. The user’s “working day” is defined as the 8 hours in which they were most active in Outlook, during two fixed months of the pre-period immediately prior to the experiment (June and July 2023).

Table 2 reports the number of unique workers and worker-week observations included in the treatment effect regressions for each outcome (defined below). ‘Unique email threads replied to’ shows the maximum observations we can include. Emails read and all session time metrics have slightly fewer work-week observations (and for emails read fewer workers included) because of an intermittent telemetry outage during the first half of 2024 that prevented us from collecting read action data for some workers in some weeks. ‘Time to reply from email delivery’ has fewer observations because it can only be defined in weeks where the worker replied to at least one email.

- **Outlook total session time** The time spent aggregated across all Outlook sessions in a week, measured in hours.
- **Outlook sessions** The number of Outlook sessions in a week.
- **Email-free work time** The total time spent during an individual’s working day in non-Outlook blocks at least 30 minutes long, measured in hours.
- **Outlook out-of-hours session time** The subset of Outlook session time that occurred outside the individual’s working day - measured in hours.
- **Unique email threads replied to** The number of unique email threads/conversations among emails that are received that week and replied to within 7 days.

¹⁹Read actions are the most commonly and frequently triggered action when people are interacting with Outlook and hence provide the best summary estimate of an Outlook session.

- **Time to reply from email delivery** The average length of time it takes to reply to an email among emails received that week and replied to within 7 days, measured in hours.
- **Emails read** The number of distinct emails read within a week.
- **Total emails in threads replied to** The total number of emails in the threads/conversations among emails that are received that week and replied to within 7 days.
- **Average email read duration** The average time spent interacting with (reading) an email, measured in seconds.
- **Number of email recipients** The average number of recipients on a worker’s sent email, observed over a fixed pre-period (April-July 2023).

B.1.2 Teams

All time metrics are measured in hours. Referring again to the last two columns of Table 2, ‘Total Teams meetings attended’ shows the maximum workers and observations for our meeting metrics. Teams data have particularly strict deletion windows (that is, Microsoft only stores individual telemetry data for a short period). Teams meeting time and recurring meeting time include the same number of workers, but fewer observations because we were only able to capture pre-period data starting later in the summer of 2023 (the start date varies by firm based on when they joined the study). We can only define the share of meetings attended early or late in weeks when workers attended at least one meeting.

- **Teams meeting time** The time spent in scheduled Teams meetings (excluding spontaneous calls) in a week.
- **Total Teams meetings attended** The number of Teams meetings attended in a week.
- **Teams out-of-hours meeting time** The subset of Teams meeting time that occurred outside the individual’s working day (see above).
- **Share of meetings attended early/late (%)** The share of scheduled Teams meetings joined either ≥ 5 minutes after the Teams meeting was started or left ≥ 5 minutes before the meeting was ended.
- **Total recurring meetings** The number of Teams meetings that were marked recurring attended in a week. We define recurring meetings as meetings that are scheduled at the same time for the same duration week over week, regardless of whether all the same people attend in each week.
- **Recurring Teams meetings time** The time spent in recurring meetings.
- **Big, short meetings** Teams meetings of length less than 1 hour with over 8 people.
- **Big, long meetings** Teams meetings of length 1 hour or more with over 8 people.
- **Small, short meetings** Teams meetings of length less than 1 hour with 8 or fewer people.
- **Small, long meetings** Teams meetings of length 1 hour or more with 8 or fewer people.
- **Number of unique people met 1-1** The total number of people met in a 1-1 meeting, observed over a fixed pre-period (April-July 2023).

Note that the cutoff of eight people for small meetings and one hour for short meetings come from the Viva Insights product, which reports these categories. Viva based these categories on research such as Rogelberg (2019), which refers to the 8-18-1800 rule suggesting that “if you are trying to solve a problem or make a decision, keep the meeting to eight people or less.”

B.1.3 Word

Time use in Word is also measured by session time but is calculated based on intentional app actions (typing or changes to a document) rather than emails read individually or in sequence. Across metrics,

a worker is a primary editor if they make the most edits on a document; a document is complete when 90% of the total edits to the document have been completed. Some documents show stray edits long after most active work is done, which may be accidental. This 90% cutoff aligns most often with other indicators of completion such as attaching to an email or printing to pdf.

All outcomes in Table 2 except ‘Word total session time’ come from an underlying telemetry source that was subject to a stricter data deletion window. We collect micro data starting in June 2023, but our first week of these constructed outcomes is the first week of July 2023. Because we measure time to complete in the week the document was completed, we leave a few “runway” weeks so we can capture the first edits to documents that were completed in July. The time to complete outcomes all have fewer observations because they are only defined in worker-weeks where the worker completed at least one document (or one solo or collaborative document). All time metrics are measured in hours.

- **Word total session time** The time spent in Word sessions, identified as blocks of actions with < 15 minute gaps.
- **Avg. document time to complete** The average time between the creation (first edit) and completion of a document, for documents completed that week and where the target worker is the primary editor.
- **Avg. collaborative document time to complete** The same as “avg. document time to complete” but only for documents with multiple collaborators.
- **Avg. non-collaborative document time to complete** The same as the “avg. document time to complete” but only for documents with no collaborators (one editor).
- **Completed documents** The number of documents completed in a week where the target worker is the primary editor.
- **Completed collaborative documents** The same as “completed documents” but only for documents with multiple collaborators.
- **Others’ documents read/edited before completion** The number of documents read or edited in a week for which the target worker is not the primary editor.
- **Word out-of-hours session time** The subset of Word session time that occurred outside the individual’s working day (see above).

B.1.4 Total out-of-hours session time

The sum of Outlook, Teams, and Word out-of-hours session time, representing a maximum of the out-of-hours activity on these apps.

B.2 Job Titles and Occupation Classification

Twenty-five of the sixty-six participating firms provided departments and/or job titles for the recruited workers. Where possible, we have used this information to impute occupation. When workers had missing job titles and departments that were not informative about work tasks (for example, a geographic region) we did not attempt to impute occupation. We classify workers into Standard Occupational Classification major groups by prompting GPT-4o (model version November 20, 2024) with unique industry and job title and/or department combinations across 5 different runs. We considered unique combinations of industry, department, and job title to be classifiable into a major occupational group if at least 3 or the 5 runs agreed. Combinations without majority agreement (accounting for 1% of workers with classifiable information) were left unclassified, resulting in a total of 2,073 workers we can categorize into major groups.

Table B.1 shows the distribution of workers across groups, and also reports the share of workers classified in each occupation group that had above-median work in the three applications we study during the pre-period. While the sample of workers with imputed occupation is too small for meaningful subgroup analyses, it confirms that the participating workers span a range of work responsibilities, but are generally white-collar knowledge workers.

B.3 Coworkers

We use telemetry data to identify the closest coworkers of each person in the experiment based on a weighted sum of their joint meetings, calls, emails, chats, and document collaborations. For each pair of workers, we calculate each metric and divide by the firm-specific average. We add these normalized numbers together for an interaction score and call someone a “coworker” if the score is greater than 8.0. (Since there are five metrics, if a pair of workers have an entirely average level of interaction, their score would be 5.) If an individual has more than 15 coworkers, we take only the 15 with the highest interaction score. The average worker has 7 coworkers. Our data use agreements prevent us from tracking individual work outcomes for the coworkers of our participating workers unless the coworkers are also study participants. However, we can aggregate across these workers to measure, for example, what share of an individual’s coworkers have a Copilot license during the study period. Figure B.1 shows the distribution of that share.

Table B.1: Occupational Classification and App Use

Group	Proportion	Above-median use of		
		Outlook	Teams	Word
Management	0.280	0.539	0.515	0.460
Business and Financial Operations	0.268	0.525	0.358	0.464
Computer and Mathematical	0.122	0.360	0.486	0.229
Sales and Related	0.110	0.581	0.240	0.358
Architecture and Engineering	0.068	0.357	0.429	0.514
Office and Administrative Support	0.053	0.486	0.294	0.284
Legal	0.018	0.789	0.105	0.500
Arts, Design, Entertainment, Sports, and Media	0.017	0.314	0.229	0.286
Life, Physical, and Social Science	0.010	0.095	0.095	0.048
Other	0.054	0.387	0.351	0.577

Note: This table reports the distribution of workers across occupation groups for the subset of workers for whom we were able to obtain meaningful job title and/or department information (2,073, or 29%), as well as the proportion of each group who had above-median pre-period use of Outlook, Teams, and Word.

Table B.2: Pre-Period Balance Test

Main Outcome Means	Control	Treated	<i>p</i>-value
Outlook total session time	11.73	11.58	0.38
Outlook sessions	31.91	31.44	0.094
Email-free work time	26.10	26.27	0.32
Outlook out-of-hours session time	2.22	2.17	0.29
Unique email threads replied to	14.41	14.60	0.57
Time to reply from email delivery	16.35	16.75	0.020
Emails read	161.37	160.32	0.73
Teams meeting time	5.17	5.28	0.26
Total Teams meetings attended	9.84	10.01	0.34
Share of mtgs. attended early/late (%)	30.36	30.92	0.19
Recurring Teams meeting time	2.28	2.35	0.20
Word total session time	1.67	1.58	0.16
Avg. document time to complete	191.11	182.14	0.20
Avg. collaborative document TTC	276.04	299.36	0.14
Avg. non-collaborative document TTC	163.61	152.75	0.13
Completed documents	0.78	0.74	0.23
Completed collaborative documents	0.14	0.13	0.22
Total out-of-hours session time	2.76	2.69	0.20
Additional Outcome Means	Control	Treated	<i>p</i>-value
Total emails in email threads replied to	3.86	3.86	0.96
Average email read duration (s)	83.77	83.70	0.97
Total recurring meetings	4.06	4.19	0.15
Big, short meetings	1.08	1.09	0.86
Big, long meetings	0.89	0.90	0.70
Small, short meetings	6.63	6.83	0.11
Small, long meetings	1.19	1.23	0.13
Teams out-of-hours meeting time	0.37	0.35	0.077
Word out-of-hours session time	0.34	0.33	0.35
Others' documents read/edited before completion	0.41	0.39	0.43
Number of email recipients (tens)	4.09	4.49	0.027
Number of unique people met 1-1 (total, tens)	13.90	14.15	0.43
Number of workers	3453	3684	—

Note: This table shows average pre-period mean weekly outcomes for control and treatment group workers. Few significant differences were found from a *t*-test on individual workers' pre-period means. All time metrics in hours. A worker is a primary editor if they make the most edits on a document. Time to complete (TTC) and completed document counts are calculated only for documents where the target worker is the primary editor.

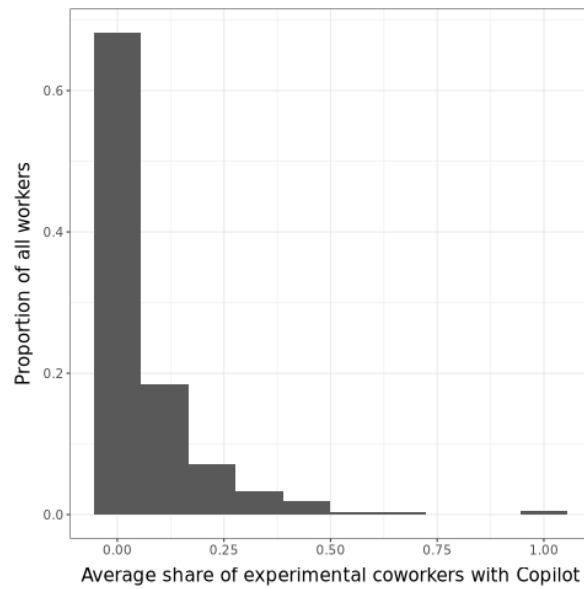


Figure B.1: Distribution of share of coworkers with Copilot

Note: This figure shows the distribution of the share of each worker’s coworkers who were in the experiment and received Copilot, averaged across post-period months 4-6. 70% of workers had on average (across weeks) less than 1 experimental coworker with Copilot, and the median worker had no experimental coworkers with Copilot. The average worker has 7 coworkers.

C Adoption

Table C.1: Copilot Usage, Overall and Across Apps

	Teams	Word	Outlook	M365 Chat	PowerPoint	Excel
Overall usage	0.661	0.542	0.493	0.421	0.253	0.086

Teams	1	0.807	0.799	0.918	0.849	0.896
Word	0.649	1	0.658	0.693	0.799	0.809
Outlook	0.608	0.625	1	0.643	0.667	0.741
M365 Chat	0.638	0.612	0.550	1	0.704	0.785
PowerPoint	0.360	0.457	0.342	0.422	1	0.568
Excel	0.127	0.152	0.129	0.160	0.193	1

Note: This table shows average weekly usage of Copilot through each integrated application among treated workers in post-period months 4-6. The first row reports overall usage rates. The following rows are conditional usage rates: given that a worker used the app Copilot in that column, what is the probability they also used the app Copilot in that row? Cells involving Teams Copilot only consider workers at Teams-using firms; cells involving Word Copilot only consider above-median Word users as in the main results.

Table C.2: Adoption Regressions with Firm-Level Average Characteristics

Dependent Variable: Model:	Individual average Copilot use								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Firm average emails read (thousands)			-0.532 (0.177)	-0.377 (0.210)					-0.573 (0.251)
Firm average unique email threads replied to (thousands)			3.67 (1.63)	4.38 (1.82)					8.33 (2.31)
Firm average Teams meeting time (hours)			0.021 (0.007)	0.006 (0.008)					0.005 (0.010)
Firm average total Teams meetings attended			0.011 (0.004)	0.013 (0.004)					0.008 (0.005)
Firm average Word total session time (hours)			0.042 (0.011)	0.041 (0.012)					0.042 (0.016)
Firm average others' documents read/edited before completion			0.040 (0.032)	0.001 (0.035)					-0.036 (0.043)
Firm average number of email recipients (tens)			0.090 (0.023)	0.098 (0.025)					0.113 (0.031)
Firm average number of unique people met 1-1 (total, tens)			-0.007 (0.015)	-0.015 (0.017)					-0.011 (0.020)
Experiment start calendar month: 2023-10-01								0.006 (0.027)	-0.031 (0.030)
Experiment start calendar month: 2023-11-01								-0.092 (0.026)	-0.070 (0.029)
Experiment start calendar month: 2023-12-01								-0.020 (0.029)	-0.024 (0.034)
Experiment start calendar month: 2024-01-01								0.161 (0.028)	0.087 (0.034)
Experiment start calendar month: 2024-02-01								-0.016 (0.038)	0.017 (0.042)
Experiment start calendar month: 2024-03-01								-0.084 (0.031)	-0.044 (0.036)
Experiment start calendar month: 2024-04-01								0.022 (0.036)	-0.034 (0.040)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes
<i>Fixed-effects</i>									
Firm					Yes	Yes			
Industry							Yes	Yes	Yes
Observations	3,422	3,422	3,684	3,422	3,684	3,422	3,684	3,684	3,422
R ²	0.092	0.092	0.087	0.120	0.181	0.213	0.029	0.081	0.151

Note: This table reports the coefficients on firm pre-period behavior and experiment start month from the regressions reported in Table 1. Coefficients on experiment start months are relative to a September 2023 baseline.

D Additional Outcomes

Table D.1: Effects of Copilot in All Post-Period Weeks

Metric	(1)	(2)	(3)	(4)	(5)
	Pre-period mean (SD)	All-months OLS (SE)	All-months IV (SE)	Workers	n
Outlook total session time	11.65 (6.81)	-1.29 (0.10)	-1.91 (0.15)	6441	313461
Outlook sessions	31.67 (11.08)	-3.58 (0.24)	-5.30 (0.36)	6441	313461
Email-free work time	26.18 (6.77)	1.44 (0.11)	2.13 (0.16)	6441	313461
Outlook out-of-hours session time	2.19 (1.84)	-0.30 (0.031)	-0.45 (0.046)	6441	313461
Emails read	160.83 (117.74)	-7.46 (1.28)	-11.08 (1.91)	6406	280932
Unique email threads replied to	14.51 (14.17)	-0.18 (0.14)	-0.26 (0.20)	6441	337149
Time to reply from email delivery	16.55 (5.34)	-0.17 (0.13)	-0.25 (0.19)	6435	297804

Teams meeting time	5.22 (3.65)	0.081 (0.046)	0.11 (0.064)	6170	221116
Total Teams meetings attended	9.93 (6.93)	0.20 (0.093)	0.28 (0.13)	6170	319836
Share of mtgs. attended early/late (%)	30.65 (11.94)	0.52 (0.23)	0.71 (0.32)	6101	186486
Recurring Teams meeting time	2.32 (1.93)	0.025 (0.025)	0.035 (0.034)	6170	268353

Word total session time	1.63 (1.58)	0.11 (0.055)	0.15 (0.075)	2525	132213
Avg. document time to complete	186.54 (94.20)	-4.93 (7.16)	-6.57 (9.53)	2525	40366
Avg. collaborative document TTC	287.85 (127.78)	-41.60 (17.38)	-52.72 (22.04)	1910	11836
Avg. non-collaborative document TTC	158.07 (89.13)	0.55 (7.21)	0.73 (9.64)	2514	34383
Completed documents	0.76 (0.80)	0.030 (0.021)	0.041 (0.029)	2525	104438
Completed collaborative documents	0.14 (0.21)	0.012 (0.0062)	0.016 (0.0084)	2525	104438

Total out-of-hours session time	2.72 (2.16)	-0.24 (0.042)	-0.35 (0.061)	6304	307005

Note: This table repeats main text Table 2 using data from every post-period week instead of only months 4-6. For easier comparison with the main specification, the IV estimates still estimate the effect of any use of Copilot in post-period months 4-6, instrumented with treatment assignment.

Table D.2: Effects of Copilot on Additional Outcome Measures

Metric	Pre-period mean (SD)	OLS (SE)	IV (SE)	Workers	n
Total emails in email threads replied to	3.86 (0.84)	-0.019 (0.018)	-0.027 (0.026)	6432	235121
Average email read duration (s)	83.73 (64.28)	-1.82 (1.26)	-2.72 (1.89)	6387	223680

Total recurring meetings	4.13 (3.45)	0.11 (0.051)	0.16 (0.071)	6170	202314
Big, short meetings	1.08 (1.26)	0.061* (0.021)	0.085* (0.030)	6170	202314
Big, long meetings	0.90 (0.78)	0.024 (0.013)	0.034 (0.019)	6170	202314
Small, short meetings	6.73 (4.63)	0.13 (0.070)	0.18 (0.098)	6170	202314
Small, long meetings	1.21 (0.98)	0.0011 (0.018)	0.0015 (0.024)	6170	202314
Teams out-of-hours meeting time	0.36 (0.50)	0.0087 (0.019)	0.012 (0.027)	5850	239620

Word out-of-hours session time	0.33 (0.39)	0.037 (0.025)	0.051 (0.034)	2459	101985

Significance: ** $q < 0.01$, * $q < 0.05$

Note: This table repeats the regressions in Table 2 for some additional outcomes. Asterisks are computed from sharpened q -values that account for the total number of hypotheses tested in the main text and this appendix table. See the notes to Table 2 for additional details.

Table D.3: Reduced Form Regressions of Outcomes on Application-Specific Copilot Usage

Metric	(1) Outlook Copilot	(2) Teams Copilot	(3) Word Copilot	(4) Other Copilot
Outlook total session time	-1.88 (0.19)	-0.33 (0.20)	-0.18 (0.19)	-0.16 (0.22)
Outlook sessions	-6.07 (0.45)	-0.35 (0.45)	0.27 (0.45)	-0.56 (0.50)
Email-free work time	2.16 (0.21)	0.25 (0.22)	0.17 (0.21)	0.26 (0.23)
Outlook out-of-hours session time	-0.45 (0.053)	-0.091 (0.057)	-0.029 (0.055)	0.0041 (0.061)
Emails read	-11.86 (2.35)	-1.85 (2.55)	-0.84 (2.31)	-0.88 (2.67)
Unique email threads replied to	0.31 (0.22)	0.16 (0.24)	0.026 (0.23)	-0.14 (0.25)
Time to reply from email delivery	-0.0061 (0.22)	-0.56 (0.26)	-0.38 (0.24)	0.23 (0.26)

Teams meeting time	0.038 (0.076)	0.061 (0.086)	0.14 (0.080)	0.10 (0.088)
Total Teams meetings attended	0.17 (0.14)	0.43 (0.16)	0.18 (0.15)	0.25 (0.17)
Share of mtgs. attended early/late (%)	0.23 (0.37)	0.010 (0.42)	0.39 (0.37)	-0.13 (0.42)
Recurring Teams meeting time	0.019 (0.038)	0.086 (0.044)	0.0023 (0.040)	-0.022 (0.047)

Word total session time	-0.067 (0.090)	0.10 (0.10)	0.36 (0.10)	-0.033 (0.10)
Avg. document time to complete	-19.19 (11.64)	9.03 (13.26)	-21.03 (12.85)	12.41 (13.37)
Avg. collaborative document TTC	-56.64 (31.26)	-11.36 (33.60)	-29.57 (32.87)	34.57 (33.22)
Avg. non-collaborative document TTC	-1.04 (11.84)	14.34 (13.60)	-29.08 (13.99)	7.58 (14.14)
Completed documents	-0.063 (0.038)	0.049 (0.039)	0.12 (0.037)	0.012 (0.040)
Completed collaborative documents	-0.024 (0.011)	0.0087 (0.011)	0.023 (0.010)	0.016 (0.011)

Total out-of-hours session time	-0.37 (0.070)	0.0011 (0.077)	0.28 (0.076)	-0.13 (0.081)

Note: This table shows estimates resulting from reduced-form regressions of our main outcomes on weekly indicators for any use of Copilot through each of our studied applications (pooling the remaining applications) These results include all fixed effect controls described in Equation 1. Note that the treatment in these regressions is endogenous (Copilot use, not randomized access) and we cannot estimate a LATE because we have only one instrument (treatment assignment) for multiple endogenous regressors. We still control for some confounding variation through the difference-in-difference framework, but nonetheless interpret these estimates as suggestive correlations rather than causal effects.

Table D.4: Treatment Effect Heterogeneity by Pre-Period Behavior

(a) Best Linear Predictors of Conditional Average Treatment Effects (CATEs) by Proxy CATEs

	Elastic Net		Neural Network	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Outlook total session time	-1.364	0.282	-1.339	0.043
	(-1.884, -0.846)	(-0.518, 1.110)	(-1.858, -0.821)	(-0.136, 0.229)
	[0.000]	[0.440]	[0.000]	[0.626]

(b) Group-Average Treatment Effects by Proxy CATEs

	Elastic Net			Nnet		
	20% Most	20% Least	Difference	20% Most (G_5)	20% Least (G_1)	Difference
Outlook total session time	-1.649	-1.177	0.532	-1.467	-1.029	0.466
	(-3.090, -0.251)	(-2.24, -0.139)	(-1.187, 2.255)	(-2.741, -0.213)	(-2.165, 0.106)	(-1.279, 2.197)
	[0.021]	[0.027]	[0.551]	[0.022]	[0.075]	[0.505]

Note: These tables show estimates of heterogeneous effects of Copilot on Outlook total session time using predictive learners from Chernozhukov et al. (2025). Elastic net and neural net methods were used to create proxy treatment effect estimates conditional on pre-period behaviors from Table 1. The first panel shows estimates of the best linear predictor of the true conditional average treatment effects (CATEs) by the proxy estimates. The second panel shows estimates of the true CATEs, grouping workers by quintile of their proxy estimate. All statistics shown are median estimates among 30 sample splits.

Table D.5: Effects of Copilot - Split by Coworkers with Copilot

Outcome	Pre-period mean (SD)		OLS (SE)		IV (SE)		n	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Below	Above	Below	Δ (Above)	Below	Δ (Above)	Below	Above
Outlook total session time	11.40 (7.22)	11.88 (6.42)	-1.30** (0.19)	-0.17 (0.26)	-1.70** (0.24)	-0.69* (0.26)	115641	134504
Outlook sessions	30.65 (12.13)	32.57 (10.03)	-3.25** (0.40)	-0.68 (0.59)	-4.42** (0.52)	-1.82* (0.61)	115641	134504
Email-free work time	26.61 (7.28)	25.81 (6.26)	1.34** (0.20)	0.32 (0.28)	1.77** (0.25)	0.88* (0.28)	115641	134504
Outlook out-of-hours session time	2.17 (1.91)	2.22 (1.77)	-0.33** (0.058)	-0.0062 (0.077)	-0.42** (0.070)	-0.13 (0.074)	115641	134504
Emails read	163.63 (129.55)	158.39 (107.73)	-6.13 (2.50)	-1.50 (3.27)	-7.62* (2.93)	-5.03 (3.07)	102415	121287
Unique email threads replied to	14.09 (14.26)	14.87 (14.03)	-0.020 (0.25)	-0.24 (0.34)	0.16 (0.29)	-0.79* (0.31)	123711	144176
Time to reply from email delivery	16.31 (5.44)	16.77 (5.26)	-0.43 (0.25)	0.20 (0.34)	-0.41 (0.29)	-0.12 (0.30)	105865	129256
Teams meeting time	4.64 (3.71)	5.77 (3.52)	0.18 (0.077)	-0.14 (0.11)	0.27* (0.090)	-0.25* (0.096)	73069	81934
Total Teams meetings attended	8.82 (7.18)	10.96 (6.55)	0.37 (0.16)	-0.28 (0.22)	0.81** (0.18)	-0.97** (0.19)	120781	132942
Share of mtgs. attended early/late (%)	32.57 (13.30)	28.90 (10.70)	0.062 (0.43)	-0.18 (0.55)	0.13 (0.45)	-0.19 (0.45)	58198	70462
Recurring Teams meeting time	2.12 (2.01)	2.51 (1.87)	0.092 (0.041)	-0.11 (0.058)	0.13* (0.047)	-0.15* (0.050)	93202	109112
Word total session time	1.66 (1.62)	1.60 (1.54)	0.15 (0.099)	-0.015 (0.13)	0.23 (0.11)	-0.067 (0.11)	47668	57245
Avg. document time to complete	181.22 (85.21)	191.12 (99.39)	-8.87 (13.19)	10.49 (18.00)	-4.03 (13.43)	-2.42 (14.55)	12990	16693
Avg. collaborative document TTC	278.47 (105.44)	294.27 (138.76)	-10.54 (34.91)	-62.68 (45.41)	-32.10 (33.41)	-48.48 (36.35)	3349	5351
Avg. non-collaborative document TTC	157.59 (91.69)	158.48 (87.37)	-10.47 (13.27)	22.07 (18.31)	-2.45 (13.46)	5.40 (15.01)	11297	13955
Completed documents	0.75 (0.83)	0.77 (0.78)	0.076 (0.037)	-0.096 (0.051)	0.050 (0.042)	-0.021 (0.046)	34743	42395
Completed collaborative documents	0.12 (0.21)	0.15 (0.21)	0.028 (0.010)	-0.032 (0.014)	0.016 (0.012)	-0.0028 (0.013)	34743	42395
Total out-of-hours session time	2.66 (2.22)	2.78 (2.11)	-0.25* (0.077)	0.0029 (0.10)	-0.38** (0.091)	0.042 (0.097)	112474	132470

Significance: ** $q < 0.01$, * $q < 0.05$

Note: This table replicates the analysis in Table 2, comparing estimated effects for a baseline group of workers who had a below-median number of close coworkers with Copilot, averaged over the post-period, to estimates for the corresponding above-median group. Significant values in Columns 4 and 6 therefore indicate significant differences in estimated effects. Standard errors are clustered by individual and q -values are calculated using adjustments for all hypotheses tested.

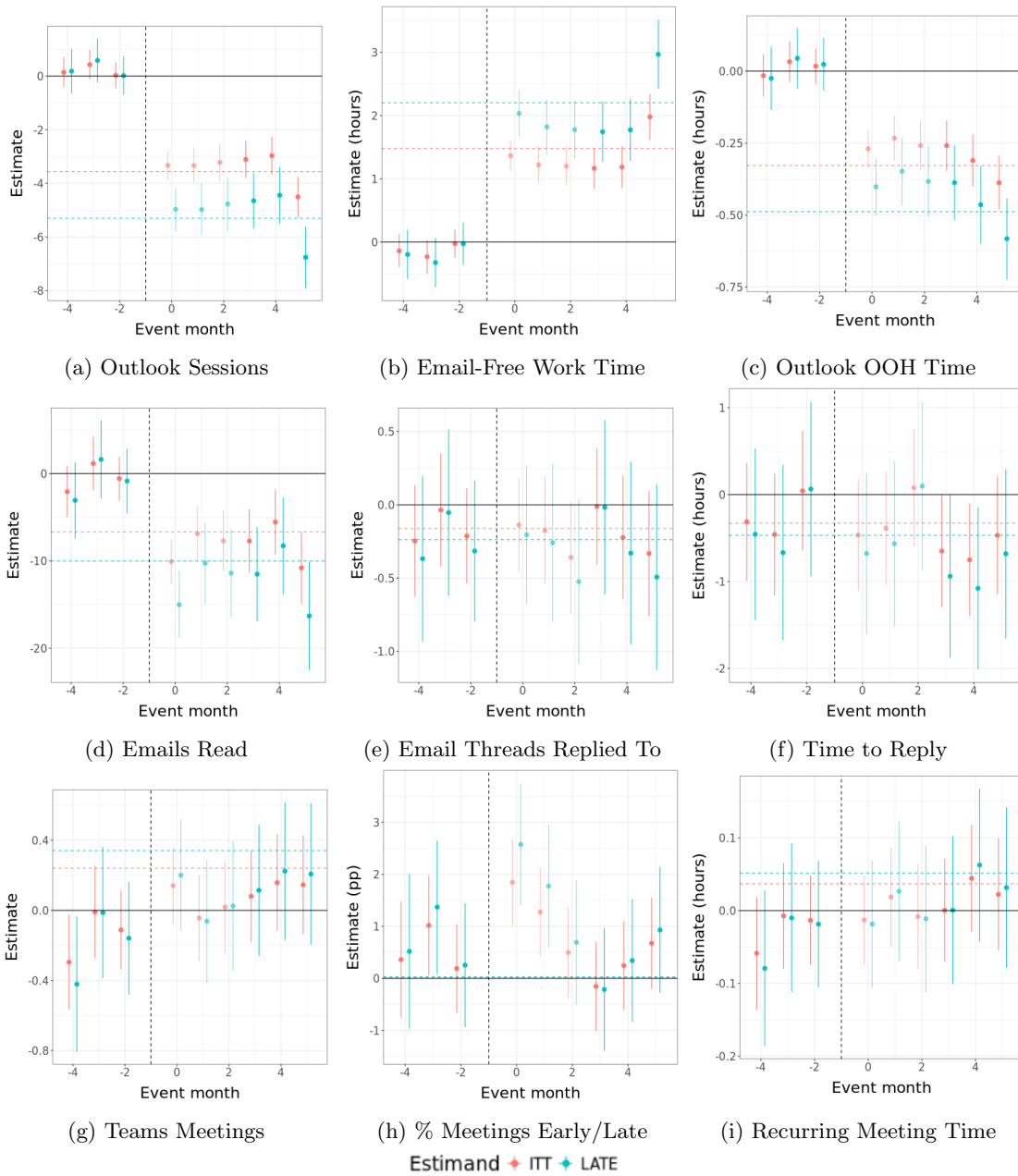


Figure D.1: Event Study Plots for the Effect of Copilot on Other Outcomes

Note: These plots show the coefficients from regressions similar to Equation (1) for the ITT (intent to treat) estimates and Equation (2) for the LATE estimates. Vertical lines show 95% confidence intervals based on individual-clustered standard errors. Horizontal dashed lines give the average estimated effect from Table 2. We include the coefficients for event months 0-2 for context.

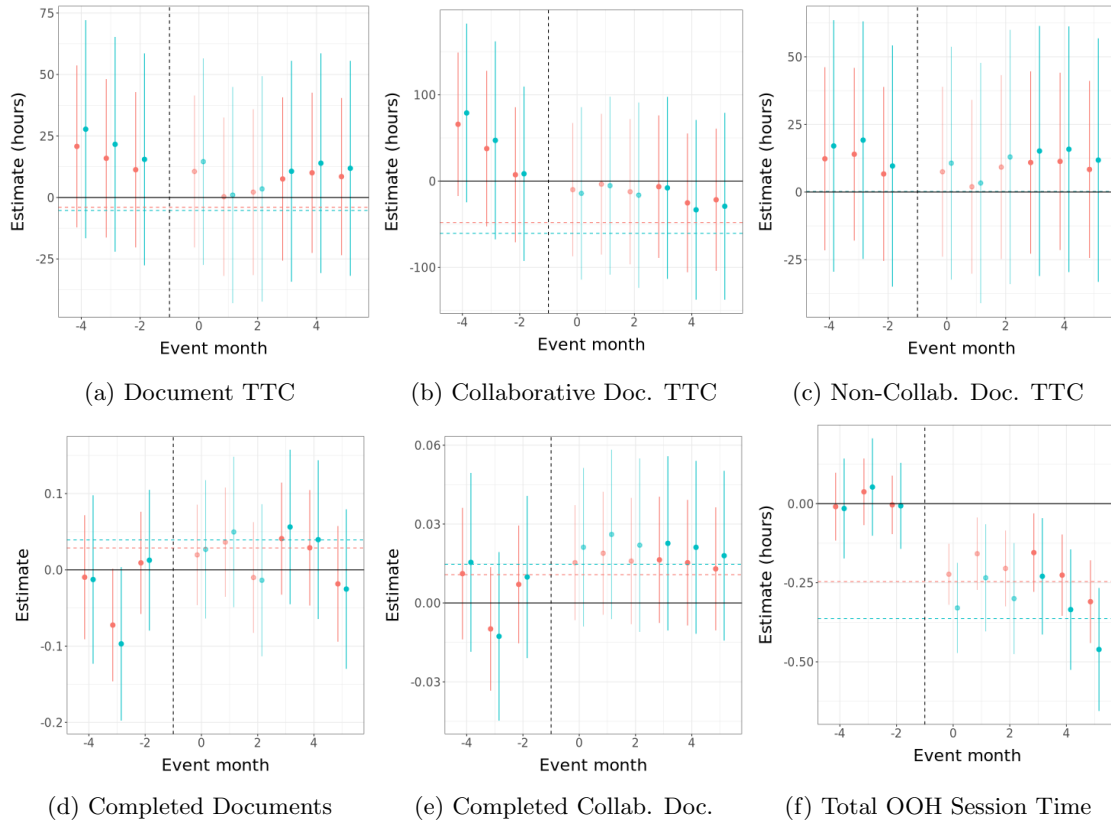


Figure D.2: Event Study Plots for the Effect of Copilot on Other Outcomes - Continued

Note: These plots show the coefficients from regressions similar to Equation (1) for the ITT (intent to treat) estimates and Equation (2) for the LATE estimates. Vertical lines show 95% confidence intervals based on individual-clustered standard errors. Horizontal dashed lines give the average estimated effect from Table 2. We include the coefficients for event months 0-2 for context.

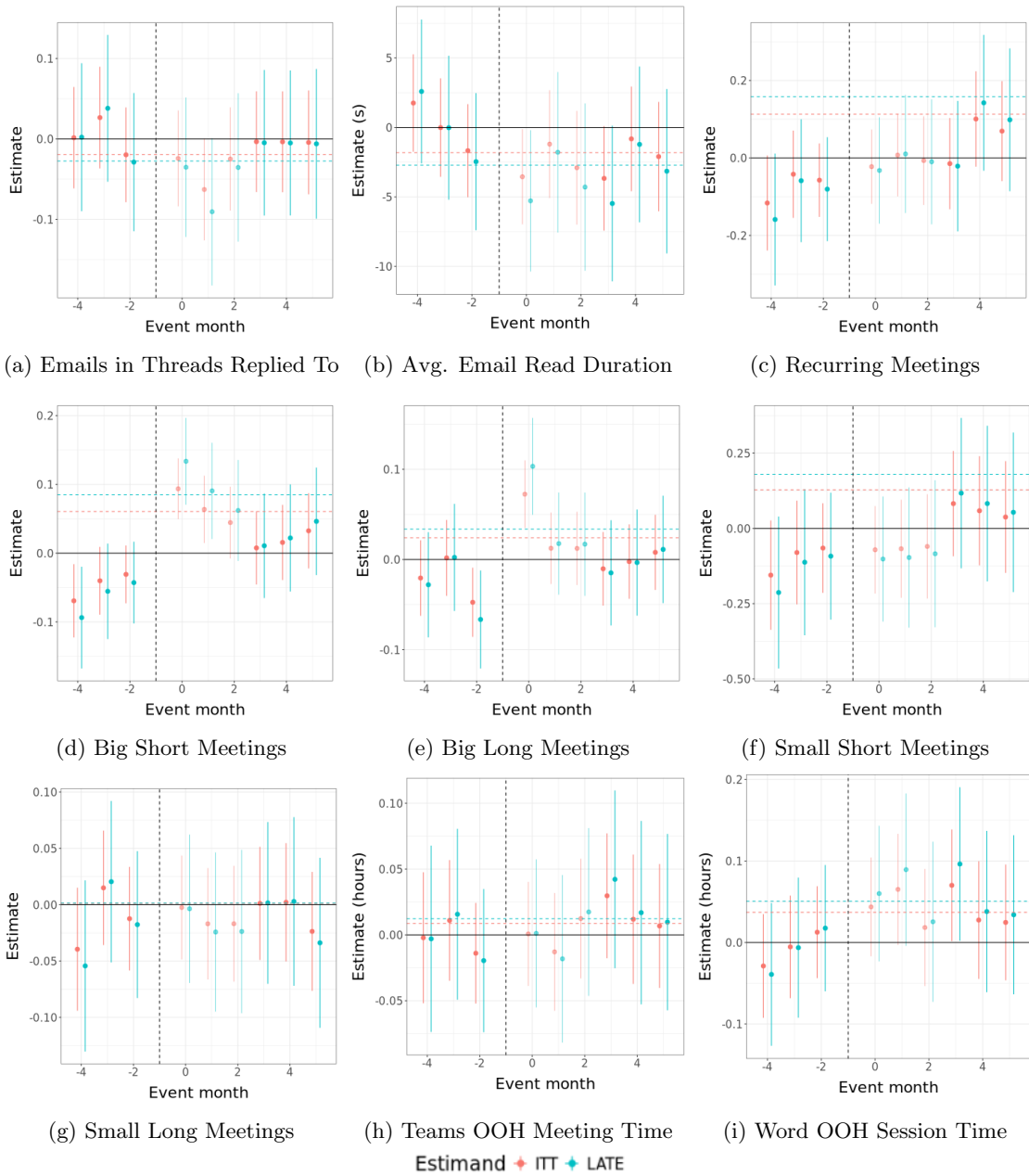


Figure D.3: Event Study Plots for the Effect of Copilot on Table D.2 Outcomes

Note: These plots show the coefficients from regressions similar to Equation (1) for the ITT (intent to treat) estimates and Equation (2) for the LATE estimates. Vertical lines show 95% confidence intervals based on worker-clustered standard errors. Horizontal dashed lines give the average estimated effect from Table D.2. We include the coefficients for event months 0-2 for context.