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 May 2025

We thank the Microsoft Customer Research Program, especially Alexia Cambon, Sida Peng, 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, and Rem Koning, Anders Humlum, and Danielle Li for helpful comments. Dillon, Jaffe, and Immorlica were employed by Microsoft, the maker of M365 Copilot, at the time of the study. Microsoft reviewed this paper for privacy concerns, but authors had full discretion over the results and estimates presented. Stanton has no relevant financial relationships related to the research. 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

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 May 2025
JEL No. L23, M1, M15, M5, O33

ABSTRACT

We present evidence on how generative AI changes the work patterns of knowledge workers using data from a 6-month-long, cross-industry, randomized field experiment. Half of the 7,137 workers in the study received access to a generative AI tool integrated into the applications they already used for emails, document creation, and meetings. We find that access to the AI tool during the first year of its release primarily impacted behaviors that workers could change independently and not behaviors that require coordination to change: workers who used the tool in more than half of the sample weeks spent 3.6 fewer hours, or 31% less time on email each week (intent to treat estimate is 1.3 hours) and completed documents moderately faster, but did not significantly change time spent in meetings.

Eleanor W. Dillon
Microsoft Research
eldillon@microsoft.com

Sonia Jaffe
Microsoft Research
spj@uchicago.edu

Nicole Immorlica
Microsoft Research
nicimm@gmail.com

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

May 6, 2025

Abstract

We present evidence on how generative AI changes the work patterns of knowledge workers using data from a 6-month-long, cross-industry, randomized field experiment. Half of the 7,137 workers in the study received access to a generative AI tool integrated into the applications they already used for emails, document creation, and meetings. We find that access to the AI tool during the first year of its release primarily impacted behaviors that workers could change independently and not behaviors that require coordination to change: workers who used the tool in more than half of the sample weeks spent 3.6 fewer hours, or 31% less time on email each week (intent to treat estimate is 1.3 hours) and completed documents moderately faster, but did not significantly change time spent in meetings.

Recent advances in generative AI have opened new possibilities for using this technology to assist with or automate a variety of tasks. Early studies have already shown that these AI can substantially increase workers' productivity in targeted tasks.[1] Over time, these advances have the potential to shift the nature of jobs (Eloundou et al., 2024) and perhaps the macroeconomy (Autor (2024), Acemoglu (2025)). However, the transition from the creation of new technologies to effective, widespread adoption in the workplace takes time (Brynjolfsson, Rock and Syverson, 2021). In lab experiments, participants are asked to focus on completing a single task and directed to use AI tools that are well-suited to that task. Outside the lab, workers must learn on their own which parts of their job benefit from the use of new tools and gradually change their habits of work. Larger productivity gains often require organizational changes to redesign processes and shift responsibilities across workers (Bresnahan, Brynjolfsson and Hitt, 2002).

In this paper, we present early evidence on how generative AI tools are changing 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 workers for the study; workers were then randomly assigned to either receive access to this new generative AI tool or

*Corresponding author: eldillon@microsoft.com. † denotes equal contribution. We thank the Microsoft Customer Research Program, especially Alexia Cambon, Sida Peng, 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, and Rem Koning, Anders Humlum, and Danielle Li for helpful comments.

[1] See, e.g., Brynjolfsson, Li and Raymond (2025), Peng et al. (2023), Noy and Zhang (2023), Dell'Acqua et al. (2023).

continue working with their current technologies. Participants in the study span many occupations, but all participating workers do some kind of work that involves heavy use of the Microsoft Office tools into which Copilot was integrated (e.g., emailing, video meetings, chat, document creation). Firms were asked to maintain the random assignment for six months, giving treated workers sufficient time to explore the new tool and integrate it into their daily work.

The study took place during the early rollout of Copilot, when firms were restricted to purchasing a small number of licenses.[2] This particular generative AI tool was new for all workers, and the median worker has only one close coworker with access to the tool (11.9% of their team). Treated workers were free to use or not use the tool during their regular work. Treated and control workers may have also used other generative AI tools during the study (for example, ChatGPT had been publicly available for nearly a year by the beginning of the study), but access to Copilot could make it easier for workers to use generative AI at work. Copilot integrates AI into applications these workers were already using regularly in their jobs, allowing some new uses, like summarizing emails or searching for information within work content, while potentially reducing the effort cost of using generative AI to create content within documents and emails.

We use highly detailed telemetry data collected by Microsoft products to track the behavior of the over 7,000 studied workers during the experiment and for several months before the introduction of the new tool. These telemetry measures capture the actions of workers, including time spent reading emails, attending video meetings, and time from creation to completion of documents. We do not observe the content of any work.

We use the product telemetry data to both confirm that workers assigned to the treatment group received a Copilot license and to track what share of workers used Copilot tools in any given week. Patterns of Copilot usage are interesting in themselves. The usage data also allow us to estimate treatment effects that adjust for low usage or noncompliance with the randomized assignment using an instrumental variable approach. We present these estimates of the local average treatment effect (LATE) alongside intent-to-treat estimates based only on workers' assigned study arm.

Over 90% of the workers assigned to receive Copilot used it at least once in the 6 month period after treatment, while usage intensity varied widely. The average treated worker used Copilot at least once in 41% of the study weeks, with a standard deviation of 30%. Usage rates varied widely across firms. In the lowest-use firm, treated workers averaged using Copilot in only 6.3% of post-period weeks. In the highest-use firm, treated workers averaged using Copilot in 75% of the post-period weeks. This variation is consistent with other surveys of AI adoption across firms (Bonney et al., 2024; McElheran et al., 2024), but the variation is notable given our sample is all large firms that were interested in being early adopters of this tool and participating in a study of its use. Industry differences and variations in workers' roles (proxied by pre-study worker behavior) explain some of the heterogeneous usage. Nonetheless, indicators for a worker's firm remain by far the most powerful predictor of usage, suggesting an important role for firm-specific differences that we are unable to measure, such as training or managerial practices.

We find clear evidence that workers who use Copilot begin to change their work patterns in meaningful ways. During this early adoption phase, we see larger changes in behaviors that workers can adjust independently and less movement in behaviors that require coordination with colleagues. Using Copilot in at least half the post-period weeks caused workers to spend over 3.5 fewer hours per week working on emails (1.3 fewer hours in the intent-to-treat estimates). This shift represents a 31% time savings from the 11.7 hours per week that the average studied worker spent on email in the pre-period.

[2] Each company was able to purchase 300 licenses. In all cases, this represents only a small fraction of their total employees who used Microsoft's Office suite.

Copilot users also shifted when they worked on email, consolidating email work into blocks that then open up four more hours of extended focus time each week.[3] We find suggestive evidence that treated workers who adopt the tool complete the documents they author 5-25% faster. In contrast, while workers are most likely to use Copilot during video meetings, we find no changes for Copilot users in the time spent in meetings and little shift in the types of meetings attended. Similarly, we find only small and statistically insignificant increases in total time spent working in Word and the number of documents authored, indicating that workers did not shift the nature of their work or take on new responsibilities. We find few substantive differences in these patterns among workers with more close coworkers who also had access to Copilot, suggesting that larger shifts in responsibilities require time and broad institutional efforts, not just local team coordination.

This work joins a small set of early studies of generative AI in real workplaces, including Brynjolfsson, Li and Raymond (2025), Cui et al. (2024), and Otis et al. (2023). These real-world studies complement a larger body of work understanding the impact of generative AI on controlled tasks in a lab setting.[4] These earlier studies all focus on the impact of AI tools on specific tasks or narrow categories of workers. Surveys indicate rapid adoption of generative AI at work across a much broader set of workers, but cannot measure worker behavior; 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 a representative sample of all U.S. workers used generative AI at work by August 2024. In follow up work, Humlum and Vestergaard (2025b) find that this widespread use of AI chatbots at work has not had any measurable impact on worker earnings or hours of work, consistent with our mixed results on worker behavior. This study aims to bridge some of the gap 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. Our findings on individual time use clarify why workers might be eager to use these tools at work even without large impacts on earnings.

1 Treatment Assignment and Patterns of Copilot Adoption

At the time the experiment was run, firms could purchase at most 300 M365 Copilot licenses. The Microsoft marketing team recruited firms to join the experiment with the understanding that at least 50 of their licenses would be allocated at random between workers. Firms were asked to recruit at least twice the number of workers for the study as the number of Copilot licenses that would be randomly assigned. Firms gave treated workers access to Copilot, but did not require them to use it in specific work tasks.[5]

The median treated worker used Copilot in 39% of the weeks we followed them (the mean was 41%).[6] Figure 1a shows the distribution of these individual usage rates. 1.6% of treated workers used Copilot every week, while 10% never used it at all. Figure 1b plots the share of treated workers using Copilot for each week since their firm allocated the experimental licenses. Usage trailed off modestly after a peak of 55% when workers were first granted access, presumably due to a combination of people

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

[4] See, 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).

[5] Additional details on the experiment are in Appendix A.

[6] Note that the data reflect engagement with Copilot, not with generative AI tools overall. These adoption rates may not be indicative of broader or long-run trends. During this early adoption period workers likely had less familiarity with the capacities of generative AI and the tools themselves were less developed, which may have lowered usage. However, firms may have recruited workers for this study who were particularly well suited to being early and enthusiastic adopters.

wanting to try the new tool and the introductory training sessions offered by some firms. In the final two months, overall weekly usage held steady at around 37% of workers.

Figure 1b also shows usage rates over time separately for Microsoft 365 Copilot usage in three programs: Teams, Outlook, and Word. Copilot can be accessed through Outlook, Teams, Word, Excel, PowerPoint, and as a separate M365 Chat, and the "any" line in Figure 1b reflects total use across all apps. We focus our impact analysis on three applications because the early phase of prompt-based generative AI was most applicable to the tasks in these tools. Workers were most likely to use Copilot in Teams, which can capture transcripts of video meetings and analyze them to generate meeting summaries and follow-up items or answer user queries like "what was discussed in the last 10 minutes?". The next most common use was through Outlook, where Copilot can summarize email chains, review a worker's inbox and flag emails that require a response, answer questions about information contained in emails 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 the contents of documents. This Copilot had the lowest usage of the three we study, partially because not all studied workers used Word regularly in their jobs (see Appendix B). Copilots for Powerpoint and Excel had lower usage rates at the time of the study.

Some of the individual heterogeneity is also reflected in average firm usage rates, which ranged from as low as 6.3% to as high as 75%.[7] We do not observe many potentially important sources that likely drive this firm variation. For example, some firms likely provided training to workers on Copilot usage or may have altered managerial practices. However, we can look at how much of the variation in adoption is explained by (the proxies we have for) the type of work done at different firms (firm industry), the type of work done by study participants (pre-experiment individual work patterns[8]), and how firms distributed their Copilot licenses (the share of a worker's coworkers who have a license). We cannot estimate the causal effect of any of these factors, only the extent to which they explain the observed variation.

Table 1: Predicting Adoption

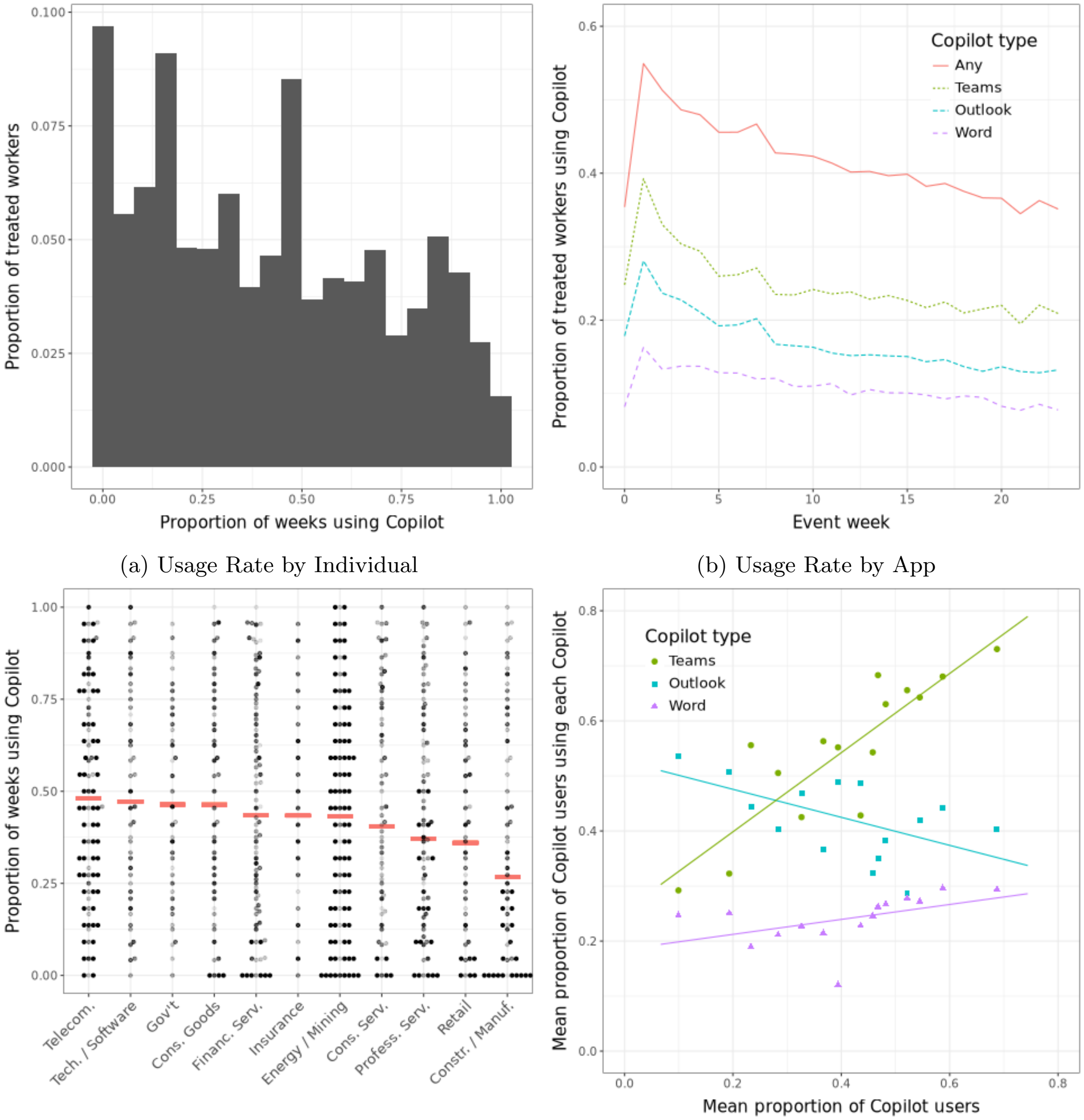
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Individ. pre-experiment behavior	$\langle \textit{all} \rangle$	$\langle \textit{all} \rangle$		$\langle \textit{all} \rangle$			$\langle \textit{all} \rangle$
Share of coworkers with Copilot		0.030 (0.040)		0.055** (0.021)			0.035 (0.037)
Firm FEs			Yes	Yes			
Industry FEs					Yes	Yes	Yes
Start-month FEs						Yes	
R^2	0.118	0.119	0.256	0.295	0.045	0.115	0.154
Within R^2				0.050			0.111

Significance: $^{**}p < 0.01$, $^{*}p < 0.05$

Note: This table reports OLS regressions where the dependent variable is the share of post-period weeks that an individual used Copilot. Because 91% of treated individuals used Copilot at least once, most of the variation is on the intensive margin. The $\langle \textit{all} \rangle$ row indicates the regression has pre-experiment work behavior included, with coefficients reported in Appendix Table A2. Standard errors are clustered by firm.

[7] Appendix Figure A3a shows how usage varied over time across firms.

[8] Weekly averages of: emails read, unique email threads replied to, Teams meeting time, Teams meetings attended, time in Word, documents edited or read before completion, number of unique email recipients, number of unique people met 1-on-1.



(a) Usage Rate by Individual

(b) Usage Rate by App

(c) Usage Rate by Industry

(d) Usage Rates across Apps by Firm Usage

Figure 1: Copilot Usage in the Treatment Group

Note: Figures (a) and (c) show the proportion of weeks in the post-period that each worker used Copilot and show the distribution across (a) workers overall or (c) workers by industry. In Figure (c), each solid dot represents 8 workers; lighter dots represent fewer than 8 workers. Mean industry usage rates are highlighted in red. Figure (b) shows average usage rates across all treated users over time. Figure (d) is a binscatter plotting the mean proportion of firms' Copilot users who are using Copilot within Outlook, Teams, and Word against firms' overall average weekly Copilot use.

Table 1 shows the R-squared from regressions of each treated worker's average share of weeks of Copilot usage over the treatment period on various individual and firm metrics. The associated coefficients are in Appendix Table A2. Individual pre-experiment activity in Outlook, Teams, and Word explains only 11.8% of the variation. The share of coworkers that have Copilot adds little explanatory power.[9] Firm fixed effects matter more, explaining 25.6% of the variation on their own and raising the R-squared from 0.119 to 0.295 when added to the other variables. Average shares of weekly use vary meaningfully across industries (ranging from 27% in Construction and Manufacturing to 48% in Telecommunications as shown in Figure 1c). Nevertheless, industry fixed effects explain less than a fifth as much of the variation in individual usage as firm fixed effects. Some of the firm variation could also come from time-of-year effects since firms joined the experiment at different times. While adding fixed effects for the start month explains some variation, it is still substantially less than the firm fixed effects.[10]

Given the importance of the firm in determining Copilot usage, we also consider whether workers at firms with high uptake rates overall are using Copilot in different applications from those in low takeup firms. Figure 1d plots the fraction of a firm's Copilot users who use Copilot within each application (averaged over weeks) against the firm's overall average Copilot usage. If patterns of use were independent of firm updake, these relationships would be flat: the same share of Copilot users would use the tool through each application regardless of their firms' overall usage. Instead, we see that workers in firms with low uptake rates are particularly likely to use Copilot through Outlook and less likely to use Copilot through Teams. These patterns suggest that Teams Copilot, and to a lesser extent Word Copilot, may be particularly valuable when multiple workers use them together. Note that while the share of Copilot users using Outlook Copilot shrinks as firm updake increases, the absolute use (the product of overall takeup and application-specific use) still increases.

2 Effects of Generative AI on Patterns of Work

2.1 Methods

Since license assignment was randomized, comparing mean treatment and control group outcomes in the post-period would give an unbiased estimate. However, since our outcome metrics vary considerably across workers and over time, a difference-in-differences (DiD) approach that exploits our pre-Copilot observations greatly improves precision. Our intent-to-treat (OLS) specification estimates the effect of receiving a Copilot license on an outcome Y_{it} for worker i in week t at firm f using an event study design:

$$Y_{it} = \alpha_i + \delta_{m_t} + \gamma_{f,\tau_{ft}} + \sum_{k \neq -1} \beta_\tau Z_i \cdot \mathbf{1}\{\tau_{ft} = k\} + \epsilon_{it} \tag{1}$$

where δ_{m_t} captures calendar year-month effects, τ_{ft} is the event month for firm f (month relative to when the firm started the experiment), and Z_i is the worker's experimental assignment. We include both calendar time and event time controls because firms entered and exited the study on a rolling basis. Note that while outcomes are measured on a weekly basis, we capture time-varying effects at a monthly frequency to reduce noise. We cluster standard errors at the firm level.

[9] The implied effect on usage is also quite small. The larger .055 coefficient implies that moving from no coworkers to the median of 11.9% would increase a person's predicted share of Copilot usage weeks by 0.65 percentage points.

[10] In the Appendix we repeat this analysis with user-week level data where we can control for calendar time. The R-squared's are overall much lower, since we are not averaging within a user, but firm fixed effects still add almost twice as much explanatory power as individual pre-experiment work patterns.

Because of our experimental design, we can define time since "treatment" for every worker in the sample, both treated workers and their comparison control workers. The effects of Copilot are identified only by differences in the month-to-month within-worker changes in outcomes between the treated workers and the comparison workers at the same firm. The potential biases from a rolling DiD analysis highlighted by (Goodman-Bacon, 2021) and others are therefore not a concern in this setup. We also use a two-way fixed effect framework where a single β on $Z_i\mathbf{1}(\tau_{ft} \geq 0)$ estimates the average treatment effect across all months of the post-period. This specification includes the same set of fixed effects, so the identifying variation is the same.

As highlighted in the previous section, over 90% of workers who had access to Copilot used it at least once. The ITT estimates are thus very close to an IV estimate where any use of Copilot over 6 months is the measure of takeup or adoption. However, many treated workers did not use Copilot regularly. In addition, some workers who were assigned to the control group nonetheless obtained Copilot licenses. To understand patterns of adoption that reflect more habitual and intense usage, we present an instrumented difference-in-difference specification (Duflo, 2001; Hudson, Hull and Liebersohn, 2017). In this case, we use treatment assignment as an instrument for whether a worker used Copilot in at least 50% of the post-treatment weeks, which allows us to understand effects for compliers who are induced to use Copilots regularly.[11] The outcome equation and first stage are, respectively:

$$\begin{aligned} Y_{it} &= \alpha_i + \delta_{m_t} + \gamma_{f,\tau_{ft}} + \beta D_{it} + \epsilon_{it} \\ D_{it} &= \tilde{\alpha}_i + \tilde{\delta}_{m_t} + \tilde{\gamma}_{f,\tau_{ft}} + \tilde{\beta} Z_i \cdot \mathbf{1}\{\tau_{ft} \geq 0\} + \eta_{it}. \end{aligned} \tag{2}$$

These estimates capture the local average treatment effect of Copilot for the complying workers who actually use it. To extend this IV setup to the multi-period event studies, we interact D_{it} with event month and instrument with interactions of Z_i and event month.

Because we test multiple hypotheses, we employ 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 similarly to a p-value after implementing a correction that accounts for the total number of hypotheses tested in the analysis.

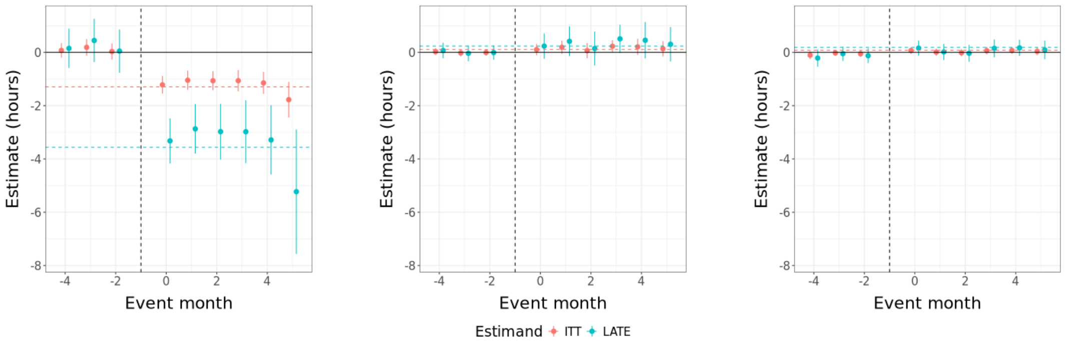
2.2 Average Effects

We are able to measure two broad sets of outcome measures: time use in digital applications and quantity-related metrics, like the number of meetings attended, documents created, etc. The quantity measures are straightforward. The time use measures for Outlook and Word are constructed from timestamps associated with user actions. Specifically, we measure time use in Outlook and Word using the concept of a session, which captures a block of time associated with an application's use. For example, an Outlook session begins when a user opens an email in the reading pane. The session ends when either: a) the user exits the reading pane and does not open another email for 15 minutes, or b) 15 minutes elapses and no other email has been opened. This stopping rule is applied to subsequent emails after a session begins. A similar concept is applied to calculate session time in Word, where sessions are based on 15 minute gaps in activities (e.g., interactions with a document or series of documents).[12]

[11] Note that D_{it} does not vary (for an individual) across weeks within either the pre- or post- period. It is one for all post-period weeks for individuals who were above-50% Copilot users and zero otherwise.

[12] Details on the construction of all worker behavior measures are in Appendix B.1.

During this early adoption phase, the most prominent impact of Copilot on worker behavior involved changes in how workers managed their email. In the pre-period, the average worker in our sample spent 11.7 hours per week reading and responding to emails, more than a quarter of a typical work week. For most workers, reading and writing emails does not represent their core job responsibilities but rather a communication overhead cost that supports their productive work. On average, workers with access to Copilot spent 1.3 hours less per week working on email in Outlook and workers who used Copilot regularly spent 3.6 fewer hours per week, a 31% decline from their pre-period average. Figure 2 illustrates that this reduction in Outlook time is apparent even in the first month after workers got access to the tool and persists throughout the study.



(a) Time in Outlook

(b) Time in Word

(c) Time in Teams Meetings

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

Note: These plots show the coefficients from regressions similar to Equations (1) for the ITT (intent to treat) estimates and Equation (2) for the LATE estimates, where the β's are allowed to vary by firm's month from Copilot adoption. Vertical lines show 95% confidence intervals based on firm-clustered errors. Horizontal dashed lines give the average estimated effect from Table 2.

Figure 2 also illustrates that we find no similar shifts in the time workers spend in Teams meetings or working in Word, despite the fact that these tasks are also complementary with the strengths of generative AI and, particularly in the case of Teams meetings, workers are regularly using Copilot through these applications. We can think of two hypothetical explanations for this difference. The first considers the demand for these tasks. When workers become more productive at a task, the effect on their total time spent doing it is ambiguous. Each unit of production in this task is now cheaper (in time) to produce, so workers may substitute towards it. For example, as generative AI improves workers' ability to capture and summarize the output of meetings, workers might choose to do more work such as brainstorming or outlining documents live in a meeting rather than asynchronously. Alternatively, if there is no demand for more of this task, workers will complete the same work as before in less total time. For example, we might see meeting times shorten or recurring meetings repeat less frequently. It could be the case that the second effect dominates in the case of email management, but the two effects roughly balance out for video meetings and document writing.

The second explanation concerns coordination across teams and organizations. During this study, the treated workers had access to Copilot, but very few of their colleagues did. In 2023, almost no companies had developed, let alone implemented, systemic plans of how generative AI could transform their workflows and allocations of responsibilities across workers. While emailing involves communicating with others, the tasks of reading emails, writing emails, and tracking conversations that require follow-ups are entirely solitary. Each worker is free to develop her own process for managing her inbox.

In contrast, meetings are group activities. Shifting what work gets done in meetings or how long they last requires coordinating with colleagues and agreeing on new norms. Writing one's own contributions to a document is a solitary process, but taking on the responsibility of being the primary author of more documents or shifting work norms so that more documents are produced requires larger changes in organizational expectations.

Table 2: Effects of Copilot Provision and Takeup

Metric	**Pre-period mean (SD)**	**OLS (SE)**	**IV (SE)**	**Workers**	**n**
Outlook total session time	11.65	-1.29**	-3.56**	6,441	313,461
	(6.81)	(0.19)	(0.44)		
Total concentration time	26.18	1.44**	3.96**	6,441	313,461
	(6.77)	(0.21)	(0.52)		
Outlook out-of-hours session time	2.19	-0.30**	-0.83**	6,441	313,461
	(1.84)	(0.057)	(0.13)		
Unique email threads replied to	13.56	-0.095	-0.24	7,137	375,642
	(13.92)	(0.12)	(0.30)		
Time to reply from email delivery	16.43	-0.19	-0.46	7,095	320,811
	(5.34)	(0.12)	(0.29)		
Teams meeting time	5.22	0.081	0.19	6,170	221,116
	(3.65)	(0.046)	(0.11)		
Share of meetings attended late/early	0.31	0.005*	0.011	6,101	186,486
	(0.12)	(0.002)	(0.005)		
Recurring Teams meetings time	2.32	0.025	0.058	6,170	268,353
	(1.93)	(0.021)	(0.048)		
Word total session time	1.63	0.11	0.24	2,525	132,213
	(1.58)	(0.098)	(0.21)		
Avg. document time to complete	186.54	-4.93	-9.80	2,525	40,366
	(94.20)	(7.10)	(14.06)		
Avg. collaborative document TTC	287.85	-41.60*	-72.81*	1,910	11,836
	(127.78)	(15.99)	(27.86)		
Avg. non-collaborative document TTC	158.07	0.55	1.11	2,514	34,383
	(89.13)	(7.73)	(15.65)		
Completed documents	0.76	0.030	0.065	2,525	104,438
	(0.80)	(0.023)	(0.049)		
Completed collaborative documents	0.14	0.012*	0.026	2,525	104,438
	(0.21)	(0.005)	(0.011)		

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

Note: For each outcome variable, this table shows pre-period mean and cross-user standard deviation (Column 1), the intent-to-treat coefficients for the effect of Copilot (Column 2) and the LATE coefficients for more frequent Copilot use (Column 3). Cross-user standard deviations are estimates of the standard deviation of the random intercept in the model $Y_{it} = \mu_i + \epsilon_{it}$. All time metrics are in hours. 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. Standard errors are clustered by firm. Asterisks reference sharpened q-values that adjust for the false discovery rate, as in Anderson (2008).

Looking in more detail at how workers changed their behavior when using Copilot, we find more evidence for the second hypothesis: workers were more likely to adjust behaviors that could be changed unilaterally and less likely to shift behaviors that required coordination with colleagues. In the top panel of Table 2, we see that workers' responsiveness and intensity of email communication with colleagues is roughly unchanged. Workers who used Copilot regularly replied to roughly the same number of conversations per week (13.8 vs. 14.1) and, if anything, replied to emails slightly faster, though the difference is not significant. These workers simply take less time to do roughly the same quantity of email work. To illustrate some of the benefits of this shift in behavior, note that workers who use Copilot not only spend less time in Outlook each week but also condense their work into fewer sessions. This opens up almost 4 hours per week of concentration time, blocks of 30 minutes or more during the regular workday with no Outlook activity. Workers using Copilot also cut back on time spent in email outside of their regular work hours.

If the null effect on total time spent in meetings was masking offsetting shifts in how workers were using meetings, we would expect to see changes in the types of meetings workers attended after gaining access to Copilot. However, as shown in Table 2, we find no impact on time spent in recurring meetings.[13] Workers with access to Copilot are slightly more likely to join meetings late or leave early, an individual choice (the IV analysis loses statistical signficance). Teams is the most common application in which workers used Copilot, so they likely found value in using generative AI in this context, perhaps in lowering effort to retain information from meetings or providing the ability to focus on participating instead of taking notes. However, this value did not translate into any visible shift in the structure of work during our sample.

Focusing on workers who were regular users of Word in the pre-period, the bottom panel of Table 2 provides more details on how Copilot access and use affects workers' document production. We do not have the power to measure these effects very precisely because over half of study participants were not regular Word users. However, there are some suggestive patterns. Consistent with Noy and Zhang (2023), we find that workers with Copilot complete their documents more quickly, although the magnitude of the effect is far smaller than what they found in the lab, and we cannot reject the null hypothesis of no effect. Interestingly, the time savings are strongest, in both time to complete (over 3 days) and percent reductions (25%), in documents where our studied worker was the primary editor but at least one other coworker also contributed. These multi-author documents take longer to complete, 12 days on average, so there may be more scope for time savings. Workers may also feel more comfortable using generative AI to stay on top of colleagues' edits and outstanding tasks compared to drafting text in work documents. The larger effects in collaborative documents also reassure us that we are seeing some behavior shift beyond casual experimentation with a new tool. Consistent with workers changing the way they do their own work, but not taking on different responsibilities, we find no change in the number of documents completed where the studied workers are the primary editor.

2.3 Effects by Peer Copilot Access

To further explore the role of team and organizational coordination in shaping the early effects of generative AI, we consider whether the effects of Copilot access and use vary when workers have more close colleagues who also have access to Copilot. As discussed in Appendix B.2, we identify close colleagues as the coworkers our studied workers interact with most frequently through emails, meetings, and joint writing. Though we do not have random variation in the share of colleagues with Copilot, the treatment and control assignments are still random for people above (or below) the median, so we can get unbiased estimates for each subgroup.

[13] As shown in appendix table A4, we find statistically significant but small (5%) increases in the count of recurring meetings and few changes in meeting type (long meetings with few participants, short meetings with many participants, etc.), which previous research has shown correlate with meeting purpose. As shown in the event studies in Figure A5, the increased participation in large meetings is concentrated in the first month of Copilot access, which suggests these might be required training sessions in the new tool.

Table 3: Effect of Copilot - Split by Share of Coworkers with Copilot licenses

Outcome	**Pre-period mean (SD)**		**OLS (SE)**		**IV (SE)**		**n**	
	Above	Below	Above	Below	Above	Below	Above	Below
Outlook total session time	11.84	11.45	-1.43**	-1.16**	-3.69**	-3.42**	168,591	144,870
	(6.37)	(7.30)	(0.25)	(0.21)	(0.57)	(0.57)		
Total concentration time	25.83	26.59	1.66**	1.23**	4.27**	3.62**	168,591	144,870
	(6.21)	(7.34)	(0.27)	(0.22)	(0.64)	(0.64)		
Outlook out-of-hours session time	2.21	2.18	-0.32**	-0.28**	-0.82**	-0.84**	168,591	144,870
	(1.75)	(1.94)	(0.067)	(0.071)	(0.15)	(0.17)		
Unique email threads replied to	14.54	12.58	-0.21	-0.047	-0.51	-0.13	189,426	186,216
	(13.93)	(13.68)	(0.18)	(0.21)	(0.42)	(0.57)		
Time to reply from email delivery	16.67	16.18	-0.14	-0.23	-0.32	-0.60	170,030	150,781
	(5.30)	(5.38)	(0.15)	(0.20)	(0.34)	(0.53)		
Teams meeting time	5.79	4.63	0.029	0.13	0.063	0.32	116,333	104,783
	(3.52)	(3.71)	(0.069)	(0.055)	(0.15)	(0.13)		
% Teams meetings attended late/early	0.29	0.33	0.005	0.006	0.010	0.013	101,607	84,879
	(0.11)	(0.13)	(0.003)	(0.004)	(0.006)	(0.010)		
Recurring Teams meetings time	2.51	2.11	-0.018	0.067	-0.039	0.16	143,514	124,839
	(1.86)	(2.01)	(0.030)	(0.032)	(0.065)	(0.072)		
Word total session time	1.61	1.65	0.12	0.088	0.25	0.20	72,324	59,889
	(1.55)	(1.62)	(0.082)	(0.14)	(0.17)	(0.32)		
Avg. document time to complete (TTC)	191.45	180.79	-5.37	-5.61	-10.57	-11.44	22,752	17,614
	(98.98)	(86.00)	(8.45)	(10.83)	(16.73)	(21.83)		
Avg. collaborative document TTC	294.49	277.91	-54.44*	-18.22	-94.40*	-32.43	7,399	4,437
	(136.75)	(106.21)	(16.87)	(25.15)	(30.75)	(43.57)		
Avg. non-collaborative document TTC	158.62	157.43	1.94	-1.57	3.91	-3.23	19,026	15,357
	(85.60)	(94.14)	(10.76)	(11.07)	(21.63)	(22.76)		
Completed documents	0.77	0.75	0.019	0.039	0.040	0.086	57,419	47,019
	(0.81)	(0.80)	(0.032)	(0.036)	(0.066)	(0.080)		
Completed collaborative documents	0.15	0.11	0.011	0.013	0.023	0.028	57,419	47,019
	(0.22)	(0.20)	(0.008)	(0.007)	(0.016)	(0.015)		

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

Note: This table replicates the analysis in Table 2 separately for workers who had above- or below-median shares of coworkers with Copilot licenses in the post-period. See that table note for explanations and variable definitions. Standard errors are clustered by firm and q-values are calculated using adjustments for all hypotheses tested. The q-values for testing that the coefficients between the two groups are different all show non-statistically significant differences.

Table 3 splits our sample by whether our studied workers have above or below median share (11.9%) of close coworkers with access to Copilot.[14] We hypothesize that more intense treatment at the team level could spur larger shifts in coordinated work behavior, for example changing meeting behavior or reallocating primary editor responsibilities across workers. However, we find no such patterns in this sample. For collaborative documents, the decrease in time to complete is larger (and statistically signficant) for the above-median group, but the difference relative to the below-median group is not statistically significant. This lack of significant differences could suggest that firm strategies, rather than local team coordination, is the primary driver of changing practices in response to new technology. However, we note that even our "high saturation" teams have, on average, only 37% of coworkers with access to Copilot, so we may simply be unable to detect team-level effects from this experimental design.

[14] Given the average worker has 7 coworkers and any worker has at most 15 coworkers, many above-median workers have just one coworker with Copilot, the median is 2.

3 Conclusion

In order to capture the full productivity potential of new technologies, firms frequently have to undergo complex organizational transformations (Feigenbaum and Gross, 2024; Bresnahan, Brynjolfsson and Hitt, 2002). 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 some of the earliest signals of how this individual use of generative AI is beginning to shape the patterns of work for knowledge workers. We observed workers during the first year in which generative AI tools were beginning to gain broad use in firms. 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. The changing behaviors we see are consistent with workers independently exploring these new tools and saving time on individual tasks, but the individual rollout did not exhibit the organization-wide shifts we might expect as teams and organizations begin larger transformations in response to this technology. We therefore expect that the patterns we find are only the beginning of a new era of work.

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 F, Erik Brynjolfsson, and Lorin M Hitt. 2002. "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence." *The quarterly journal of economics*, 117(1): 339–376.

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."

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.*

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.*

Dillon, Eleanor Wiske, Sonia Jaffe, Sida Peng, and Alexia Cambon. 2025. "Early Impacts of M365 Copilot." Microsoft.

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.

Feigenbaum, James, and Daniel P Gross. 2024. "Answering the Call of Automation: How the Labor Market Adjusted to Mechanizing Telephone Operation." National Bureau of Economic Research Working Paper 28061.

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.*

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.

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*.

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*.

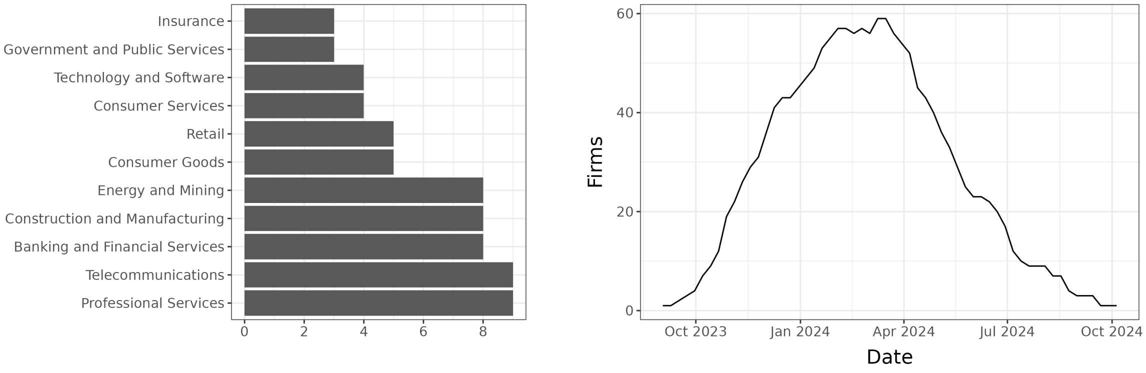
Vaithilingam, Priyan, Tianyi Zhang, and Elena L Glassman. 2022. "Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models." *Chi conference on human factors in computing systems extended abstracts*, 1–7.

Appendix

A Experiment Details

More details on the study design and implementation are available in Dillon et al. (2025). For example, a firm that chose to allocate 50 of their 300 licenses at random would need to recruit at least 100 workers to join the experiment, while a firm that chose to allocate 150 licenses at random would need to recruit 300 workers to join. Some firms had more than 500 workers participate in the study, while one firm had only 68 workers participate. Each individual firm selected which workers they wanted to participate; some focused on a single department or work area while others pulled workers from many parts of the company. Firms were responsible for all communications with the participating workers, including obtaining necessary consent. Some firms provided additional information about the recruited workers such as department or job title. While the sample of workers with this additional information is too small for meaningful subgroup analyses, it confirms that the participating workers are generally early to mid-career knowledge workers such as analysts, project managers, or professional service providers.

Our analysis sample has 66 firms across a range of industries, as shown in Figure A1a. There were 2 additional firms that 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.



(a) Industry Mix

(b) Over Time

Figure A1: 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.

For all firms, our data sample begins on May 5, 2023, providing at least 4 months of data before any Copilot licenses were allocated. Firms varied in when they allocated licenses, with the first starting in September 2023, and the latest in April 2024. Figure A1b illustrates the number of firms actively maintaining randomization over the study window. 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. 5 firms have shorter treatment periods because they explicitly exited the experiment prior to the 24-week post-period. 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.

Table A1 shows the balance tests for pre-experiment differences between the treatment and control groups. There are no significant differences across all outcomes.

Table A1: Pre-period Balance Test

Main Outcomes	**Control Mean**	**Treated Mean**	***p*-value**
Outlook total session time	11.73	11.58	0.45
Total concentration time	26.10	26.27	0.37
Outlook out-of-hours session time	2.22	2.17	0.42
Unique email threads replied to	13.52	13.59	0.84
Time to reply from email delivery	16.27	16.58	0.05
Teams meeting total time	5.17	5.28	0.23
% Teams meetings attended late/early	0.30	0.31	0.13
Recurring Teams meetings time	2.28	2.35	0.15
Word total session time	1.67	1.58	0.17
Avg. document TTC	191.11	182.14	0.26
Avg. collaborative document TTC	276.04	299.36	0.10
Avg. non-collaborative document TTC	163.61	152.75	0.14
Completed documents	0.78	0.74	0.30
Completed collaborative documents	0.14	0.13	0.28

Additional Metrics	**Control Mean**	**Treated Mean**	***p*-value**
Emails read	158.03	157.00	0.78
Total emails in email threads replied to	3.83	3.83	0.77
Total Teams meetings attended	9.84	10.01	0.40
Total recurring meetings	4.06	4.19	0.15
Big short meetings	1.08	1.09	0.87
Big long meetings	0.89	0.90	0.71
Small short meetings	6.63	6.83	0.14
Small long meetings	1.19	1.23	0.17
Others' documents read/edited	0.41	0.39	0.34
Number of email recipients	4.09	4.49	0.05
Number of unique people met 1-1 (total)	13.90	14.15	0.39

Note: This table shows average pre-period mean weekly outcomes for control and treatment group workers. No 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.

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. For email outcomes, this distinction appears unimportant. Virtually every worker in our sample used Outlook regularly (6,996 out of 7,137) in the pre-period months and we suspect we are capturing a large share of work-related email activity.

In contrast, many sample workers show almost no Word activity during our pre-period months. Based on conversations with participating firms, we think this mostly reflects roles that do not involve document creation, though some low activity could reflect use of a different word proccessing tool. We do not expect that individual access to Copilot would shift many of these workers towards using Word, but 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).

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 any 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.

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.

- **Outlook**

  For email, we aggregate the signals of various 'reading actions' individuals take into 'Outlook sessions:' blocks of time where they don't go more than 15 minutes without reading an email.[15] 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.

  - **Outlook total session time** The time spent aggregated across all sessions in a week, measured in hours.
  - **Outlook out-of-hours session time** The subset of Outlook session time that occurred outside the individual's "working day" as defined as the 8 hours in which they were most active during one fixed month of the pre-period immediately prior to the experiment.
  - **Total concentration time** The total time spent during an individual's working day in non-Outlook blocks at least 30 minutes long.
  - **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.
  - **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.
  - **Emails read** Count 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.

- **Teams**

[15] Read 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.

  - **Teams meeting time** The time spent in scheduled Teams meetings (excluding spontaneous calls) in a week, measured in hours.
  - **Share of meetings attended late/early** The share of scheduled Teams meetings joined either $\geq$5 minutes after the Teams meeting was started or left $\geq$5 minutes before the meeting was ended.
  - **Recurring Teams meetings time** The time spent in Teams meetings marked as recurring meetings, measured in hours.
  - **Total Teams meetings attended** The number of Teams meetings attended in a week.
  - **Total recurring meetings** The number of Teams meetings that were marked recurring attended in a week.
  - **Big short meetings** The number of Teams meetings of length less than 1 hour with over 8 people.
  - **Big long meetings** The number of Teams meetings of length 1 hour or more with over 8 people.
  - **Small short meetings** The number of Teams meetings of length less than 1 hour with 8 or fewer people.
  - **Small long meetings** The number of Teams meetings of length 1 hour or more with 8 or fewer people.

- **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.[16]

  - **Word total session time** The time spent in Word sessions, identified as blocks of actions with $<$ 15 minute gaps, measured in hours.
  - **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, measured in hours.
  - **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.

[16] 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.

B.2 Coworkers

We also use the 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 worker, 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. We can then calculate what share of an individual's coworkers have a Copilot license. Figure A2 shows the distribution of that share.

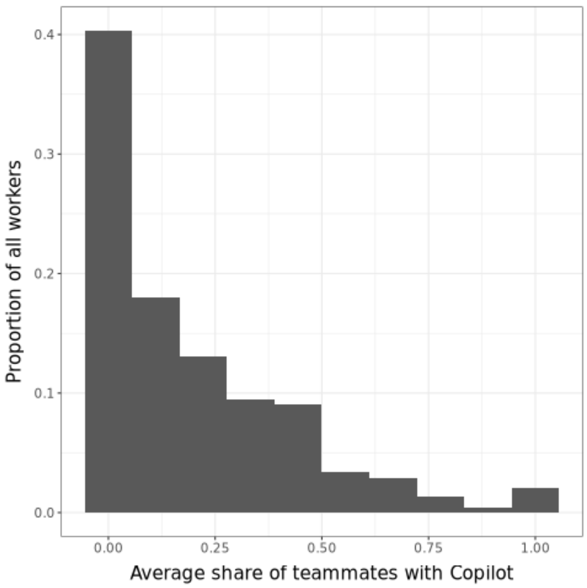


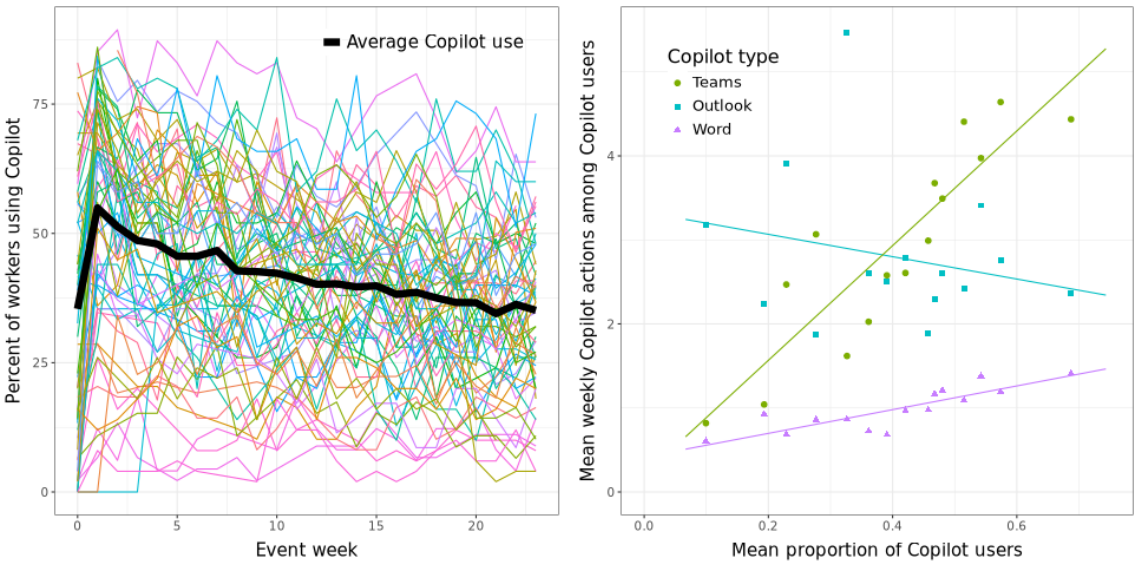
Figure A2: Distribution of share of coworkers with Copilot

C Additional Tables and Figures

C.1 Adoption

Figure A3 shows some additional Copilot usage patterns. Figure A3a shows the variation across firms in usage over time. Figure A3b shows a binscatter of the mean weekly Copilot actions taken in each app among users using some form of Copilot that week by the firm-level share of users that are active in an average week. Similar to Figure 1b, we see that the usage intensity grows with firm usage for Teams, while it is flat or even declining for Outlook (with Word in between).

Table A2 presents all the coefficients associated with the regressions in Table 1. Table A3 does the same analysis with user-week data instead of user-level averages. The R-squared's are lower since there is lots of unexplained week-to-week variation, however the pattern is similar.



(a) Usage Rate by Firm

(b) Usage Intensity across Apps by Firm Usage

Figure A3: Copilot Usage

Note: Figure (a) shows usage over weeks of the experiment, with one line for each participating firm. Figure (b) shows a binscatter of the mean weekly Copilot actions taken in each app among users using some form of Copilot that week by the firm-level share of users that are active in an average week.

Table A2: Predicting Adoption - Coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	0.3197**	0.3136**					
	(0.0273)	(0.0271)					
Emails read	-0.0003**	-0.0002**		-0.0002**			-0.0003**
	(7.69×10^{-5})	(7.4×10^{-5})		(5.58×10^{-5})			(7.3×10^{-5})
Unique email threads replied to	-0.0008	-0.0008		-0.0004			-0.0006
	(0.0007)	(0.0007)		(0.0005)			(0.0006)
Teams meeting total session time	0.0120**	0.0119**		0.0091**			0.0132**
	(0.0035)	(0.0034)		(0.0024)			(0.0031)
Teams meetings attended	0.0072**	0.0073**		0.0035*			0.0056*
	(0.0024)	(0.0024)		(0.0017)			(0.0022)
Word total session time	0.0157*	0.0161*		0.0116*			0.0202**
	(0.0070)	(0.0069)		(0.0045)			(0.0070)
Others' documents read/edited before completion	0.0422*	0.0408*		0.0325*			0.0410*
	(0.0184)	(0.0181)		(0.0138)			(0.0171)
Number of email recipients	-0.0005	-0.0005		-0.0007			-0.0007
	(0.0005)	(0.0004)		(0.0004)			(0.0004)
Number of unique people met 1-1 (total)	0.0002	0.0002		0.0003			0.0002
	(0.0006)	(0.0006)		(0.0004)			(0.0005)
Share of coworkers with Copilot license		0.0302		0.0546**			0.0352
		(0.0391)		(0.0206)			(0.0368)
Fixed-effects							
Firm			Yes	Yes			
Industry					Yes	Yes	Yes
Post-period start month						Yes	
Fit statistics							
Observations	3,422	3,422	3,684	3,422	3,684	3,684	3,422
R^2	0.11840	0.11896	0.25608	0.29475	0.04462	0.11478	0.15394
Within R^2				0.05014			0.11090

Significance: $^{**}p < 0.01$, $^{*}p < 0.05$

Note: This table shows OLS regressions results using individuals' pre-experiment behaviors, share of coworkers with a Copilot license, and firm, industry, and post-period start month fixed effects to predict mean post-period Copilot weekly usage among treated workers. All variables are averages over the pre-period other than "number of unique people met 1-1," which is a total over a fixed pre-period.

Table A3: Predicting Weekly Copilot Use

Emails read	-0.0002**	-0.0002**		-0.0002**				-0.0003**
	(7.3×10^{-5})	(6.99×10^{-5})		(5.64×10^{-5})				(6.96×10^{-5})
Unique email threads replied to	-0.0008	-0.0008		-0.0004				-0.0006
	(0.0007)	(0.0007)		(0.0005)				(0.0006)
Teams meeting total session time	0.0132**	0.0131**		0.0092**				0.0140**
	(0.0033)	(0.0032)		(0.0024)				(0.0031)
Teams meetings attended	0.0064**	0.0065**		0.0035*				0.0050*
	(0.0023)	(0.0023)		(0.0017)				(0.0022)
Word total session time	0.0158*	0.0161*		0.0115*				0.0198**
	(0.0068)	(0.0068)		(0.0046)				(0.0068)
Others' documents read/edited before completion	0.0398*	0.0389*		0.0337*				0.0393*
	(0.0176)	(0.0173)		(0.0136)				(0.0166)
Number of email recipients	-0.0005	-0.0005		-0.0007				-0.0006
	(0.0004)	(0.0004)		(0.0004)				(0.0004)
Number of unique people met 1-1 (total)	0.0003	0.0003		0.0003				0.0003
	(0.0006)	(0.0006)		(0.0004)				(0.0006)
Share of coworkers with Copilot license		0.0221		0.0535*				0.0257
		(0.0389)		(0.0204)				(0.0364)
Fixed-effects								
Calendar week	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Event month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm			Yes	Yes				
Industry						Yes	Yes	Yes
Post-period start month							Yes	
Fit statistics								
Observations	76,681	76,681	82,582	76,681	82,582	82,582	82,582	76,681
R^2	0.05984	0.05994	0.10615	0.12064	0.01791	0.03377	0.05729	0.07189
Within R^2	0.04296	0.04307		0.01517				0.03847

Significance: $^{**}p < 0.01$, $^{*}p < 0.05$

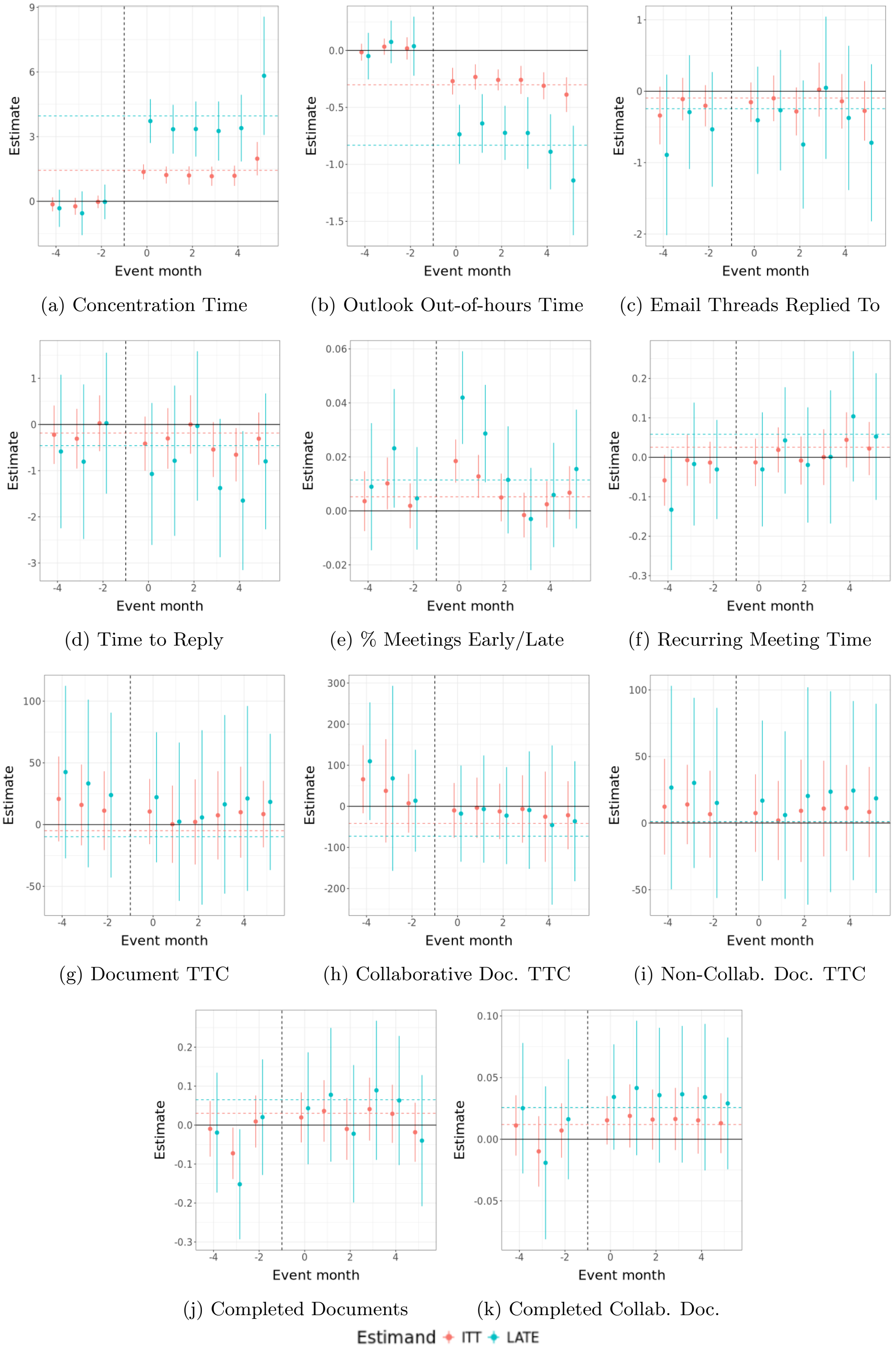
Note: This table shows OLS regressions results using individual pre-experiment behaviors, share of coworkers with a Copilot license, firm industry fixed effects, and calendar-week and event-month fixed effects to predict post-period Copilot weekly usage among treated workers. All variables are averages other than "number of unique people met 1-1," which is a total over a fixed pre-period.

C.2 Outcomes

Similar to Figure 2, Figure A4 shows event studies for the other variables in Table 2.

Table A4 shows the same analysis as Table 2 for some additional outcome variables. The decrease in emails read individually is unsurprising if Copilot is summarizing email threads (or all new emails) for people. Total emails replied to is another measure of reply activity where, while the difference is statistically significiant, it is practically very small (< 2%).

We also look at several different types of meetings. While we do see some increases for big meetings, the event studies in Figure A5 show that these increases only happened at the very start of the study period and seem likely due to the Copilot trainings that many companies conducted.



(a) Concentration Time

(b) Outlook Out-of-hours Time

(c) Email Threads Replied To

(d) Time to Reply

(e) % Meetings Early/Late

(f) Recurring Meeting Time

(g) Document TTC

(h) Collaborative Doc. TTC

(i) Non-Collab. Doc. TTC

(j) Completed Documents

(k) Completed Collab. Doc.

Figure A4: Event Study Plots for the Effect of Copilot on Other Outcomes

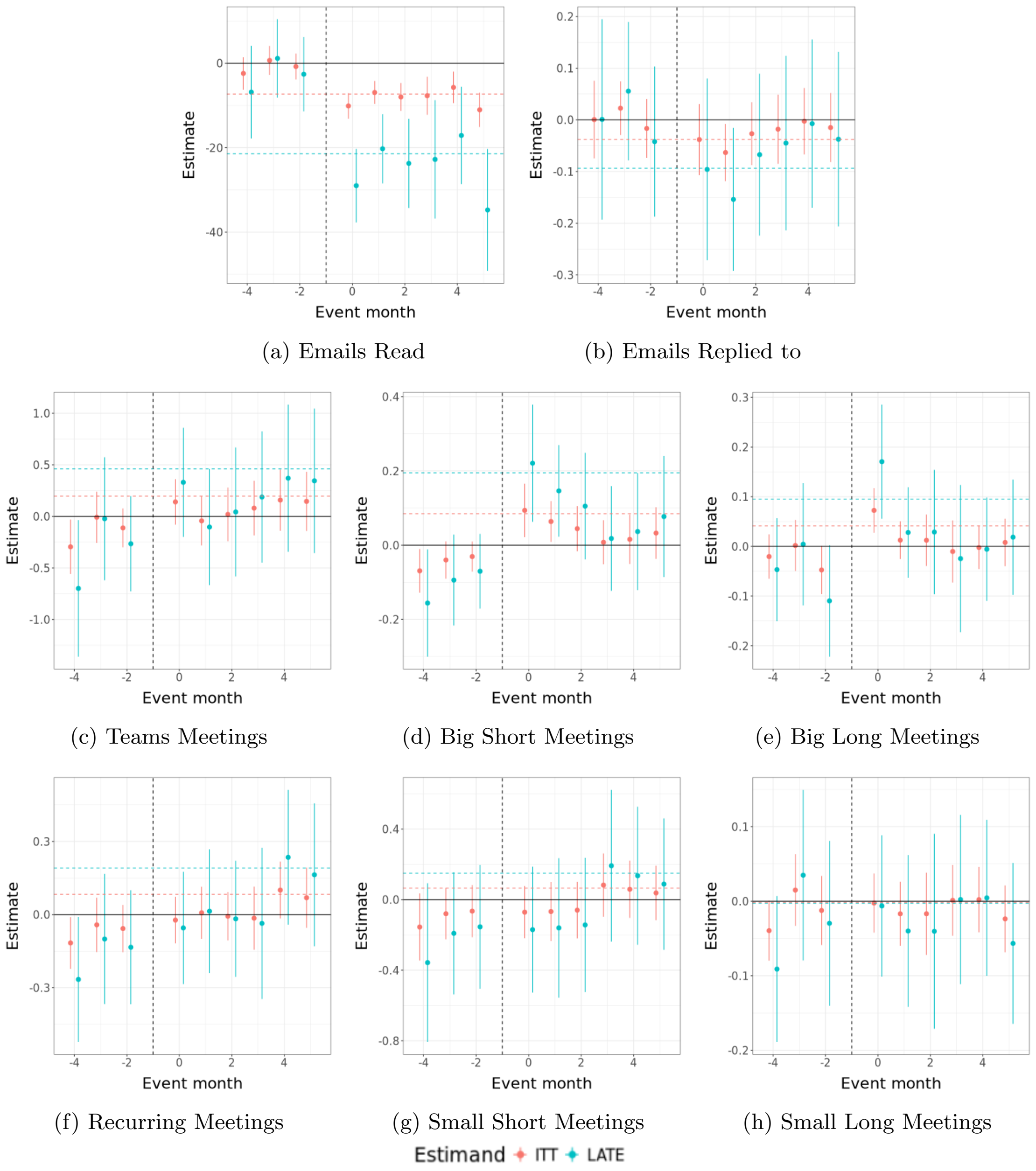
Note: These plots show the coefficients from regressions similar to Equation (1) for the ITT (intent to treat) estimates and Equation (2) for the LATE estimates. Vertical lines show 95% confidence intervals based on firm-clustered standard errors. Horizontal dashed lines give the average estimated effect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

Table A4: Effects of Copilot on Additional Outcome Measures

Metric	**Pre-period mean (SD)**	**OLS (SE)**	**IV (SE)**	**Workers**	**n**
Emails read	157.50	-7.31**	-21.43**	6,700	288,509
	(117.53)	(1.65)	(4.42)		
Total emails in email threads replied to	3.83	-0.038*	-0.093*	7,095	320,811
	(0.84)	(0.013)	(0.034)		
Total Teams meetings attended	9.93	0.20*	0.46	6,170	319,836
	(6.93)	(0.090)	(0.21)		
Total recurring meetings	4.13	0.083*	0.19*	6,170	268,353
	(3.45)	(0.033)	(0.073)		
Big short meetings	1.08	0.085**	0.19**	6,170	268,353
	(1.26)	(0.019)	(0.039)		
Big long meetings	0.90	0.041**	0.095*	6,170	268,353
	(0.78)	(0.012)	(0.028)		
Small short meetings	6.73	0.065	0.15	6,170	268,353
	(4.63)	(0.043)	(0.100)		
Small long meetings	1.21	-0.001	-0.003	6,170	268,353
	(0.98)	(0.016)	(0.036)		

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

Note: For each outcome variable, this table shows pre-period mean and cross-user standard deviation (Column 1), the intent-to-treat coefficients for the effect of Copilot (Column 2), and the LATE coefficients (Column 3). Cross-user standard deviations are estimates of the standard deviation of the random intercept in the model $Y_{it} = \mu_i + \epsilon_{it}$. Standard errors are clustered by firm, and asterisks are computed from sharpened q-values that account for the total number of hypotheses tested in the main text and appendices.



(a) Emails Read

(b) Emails Replied to

(c) Teams Meetings

(d) Big Short Meetings

(e) Big Long Meetings

(f) Recurring Meetings

(g) Small Short Meetings

(h) Small Long Meetings

Figure A5: Event Study Plots for the Effect of Copilot on Table A4 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 firm-clustered standard errors. Horizontal dashed lines give the average estimated effect from Table A4.