Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence

Shakked Noy MIT
Whitney Zhang MIT

March 10, 2023
Working Paper (not peer reviewed)

Abstract

We examine the productivity effects of a generative artificial intelligence technology—the assistive chatbot ChatGPT—in the context of mid-level professional writing tasks. In a preregistered online experiment, we assign occupation-specific, incentivized writing tasks to 444 college-educated professionals, and randomly expose half of them to ChatGPT. Our results show that ChatGPT substantially raises average productivity: time taken decreases by 0.8 SDs and output quality rises by 0.4 SDs. Inequality between workers decreases, as ChatGPT compresses the productivity distribution by benefiting low-ability workers more. ChatGPT mostly substitutes for worker effort rather than complementing worker skills, and restructures tasks towards idea-generation and editing and away from rough-drafting. Exposure to ChatGPT increases job satisfaction and self-efficacy and heightens both concern and excitement about automation technologies.

We gratefully acknowledge financial support from an Emergent Ventures grant, the George and Obie Shultz Fund, and the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. The research described in this article was approved by the MIT Committee on the Use of Humans as Experimental Subjects, and was preregistered at the AEA RCT Registry (AEARCTR-0010882). We thank Daron Acemoglu, Nikhil Agarwal, David Autor, Lucas Barros, Talia Benheim, Amy Finkelstein, John Horton, Simon Jäger, Ailidh Leslie, Jackson Mejia, Ilan Noy, Liora Noy, Emily Partridge, Charlie Rafkin, Aakaash Rao, Nina Roussille, Chris Roth, Frank Schilbach, Benjamin Schoefer, Lexi Schubert, Advik Shreekumar, Shine Wu, and participants at the MIT Labor Lunch for helpful comments and conversations.

1 Introduction

Recent advances in generative artificial intelligence may have widespread implications for production and labor markets. New generative AI systems like ChatGPT or DALL-E, which can be prompted to create novel text or visual outputs from large amounts of training data, are qualitatively unlike most historical examples of automation technologies. Previous waves of automation predominantly impacted "routine" tasks consisting of explicit sequences of steps that could be easily codified and programmed into a computer (Autor and Dorn, 2013; Autor, 2015). Creative, difficult-to-codify tasks (such as writing and image generation) largely avoided automation—a pattern that scholars noted might change with the advent of the deep learning techniques that now underpin generative AI systems.

The emergence of powerful generative AI technologies reintroduces a host of classic questions in a new context. Automation technologies—by definition—perform specific tasks in place of humans. But, more broadly, they may either displace humans completely from certain occupations or complement existing human workers and increase their productivity (Acemoglu and Restrepo, 2020; Boustan et al., 2022; Kanazawa et al., 2022). Insofar as automation technologies mostly displace human workers, they can increase unemployment, while their impacts on aggregate productivity may be small or nonexistent to the degree that they mainly serve to redistribute income from workers to owners of capital (Acemoglu and Restrepo, 2018). Insofar as automation complements existing workers, it can simultaneously benefit workers, capital owners, and consumers by raising productivity and wages and lowering prices (Kleinberg et al., 2018; Hoffman et al., 2018; Agrawal et al., 2019).

For example, a potent generative writing tool like ChatGPT might entirely replace certain kinds of writers, such as grant writers or marketers, by letting companies directly automate the creation of grant applications and press releases with minimal human oversight. This might not increase the quality of the resulting written output, but would let companies save on wage costs by eliminating human labor. Alternatively, a tool like ChatGPT could substantially raise the productivity of grant writers and marketers, for example by automating relatively routine, time-consuming subcomponents of their writing tasks, such as translating ideas into an initial rough draft. In this case, demand for these services could expand, resulting in higher employment and wages as well as greater productivity for companies and cheaper products for consumers. Inequalities between workers could also be affected: inequality could decrease if lower-ability workers are helped more by ChatGPT, or increase if higher-ability workers have the skills necessary to take advantage of the new technology.

Which of these eventualities will generative AI systems bring about? The answer depends on a host of questions: how do generative AI systems affect workers' productivity in existing tasks? Do they affect productivity largely by substituting for worker effort or complementing worker skills? Do they differentially affect low- or high-ability workers, or workers with

different skill profiles? Do they affect workers' enjoyment of their work (Schwabe and Castellacci, 2020)?

This paper takes the first step towards answering these questions.[1] In an online experiment, we recruit 444 experienced, college-educated professionals and assign each to complete two occupation-specific, incentivized writing tasks. The occupations we draw on are marketers, grant writers, consultants, data analysts, human resource professionals, and managers. The tasks, which include writing press releases, short reports, analysis plans, and delicate emails, comprise 20-to 30-minute assignments designed to resemble real tasks performed in these occupations; indeed, most of our participants report completing similar tasks before and rate the assigned tasks as realistic representations of their everyday work. Participants face high-powered incentives, in the form of large bonus payments, to produce high-quality work. Quality is assessed by (blinded) experienced professionals working in the same occupations. Evaluators are asked to treat the output as if encountered in a work setting and are incentivized to grade outputs carefully. Evaluators assign an overall grade as well as separate grades for writing quality, content quality, and originality. Each piece of output is seen by three evaluators, with an average within-essay cross-evaluator correlation of 0.44.

A randomly-selected 50% of our participants—the treatment group—are instructed to sign up for ChatGPT between the first and second task, are walked through how to use it, and are told they are permitted to use it on the second task if they find it useful. The control group is instead instructed to sign up for the LaTeX editor Overleaf. This design allows us to estimate the causal effects of ChatGPT using a combination of within-person and between-person variation, and performance on the first task serves as a measure of baseline ability that enables our inequality analyses.

We collect participants' output and elicit total time taken, time taken on various subcomponents of the task, job satisfaction, self-efficacy, and beliefs about automation. We also take a snapshot of each participant's output each minute while they perform the task, to construct an objective measure of time active on the task and to detect ChatGPT usage in the control group and on the pre-treatment task.

A complete description of our experimental design, a copy of relevant survey questionnaires, and additional figures validating our central measures and extending our main results are included in the Online Appendix. Descriptive statistics about the sample, as well as balance and selective attrition tests, are available in Table 1. The attrition rate is 5% in the control group and 10% in the treatment group. Balance tests indicate that across 13 pre-treatment characteristics, the treatment and control group exhibit a small significant difference only for only two characteristic (employment status and being an HR professional). Our partly within-person design, which controls for performance on the pre-treatment task, should

[1] A nascent literature has studied applications of machine learning to *predictive* tasks, but not generative tasks.

eliminate any influence of selective attrition on our results; in the Online Appendix, we also report Lee (2009) bounds on our main results and versions of our results controlling for employment status and occupation, which confirm that our results are highly robust to selective attrition.

2 Results

2.1 Takeup of ChatGPT

In the treatment group, 92% of treated participants successfully sign up for ChatGPT,[2] and 81% choose to use it on the second task, giving it an average self-assessed usefulness score of 4.4 out of 5.

Prior to treatment, about 70% of our participants had heard of ChatGPT and 30% had used it before. Self-reported and objective measures indicate only 10-20% of the control group use ChatGPT on the tasks, meaning there is at least a 60 percentage point experimentally-induced gap in usage between our treatment and control groups on the second task. The fact that some control participants are using the tool means our estimates provide lower bounds on the effects of ChatGPT usage on productivity.

2.2 Productivity

We measure productivity as earnings per minute. Figure 1 shows that the experimental intervention increases this outcome dramatically. In the treatment group, time taken on the post-treatment task drops by 10 minutes (37%) relative to the control group, who take an average of 27 minutes ($p = 0.000$). Average evaluator grades in the treatment group increase by 0.45 standard deviations ($p = 0.000$), with roughly similar increases for overall grades and specific grades for writing quality, content quality, and originality.

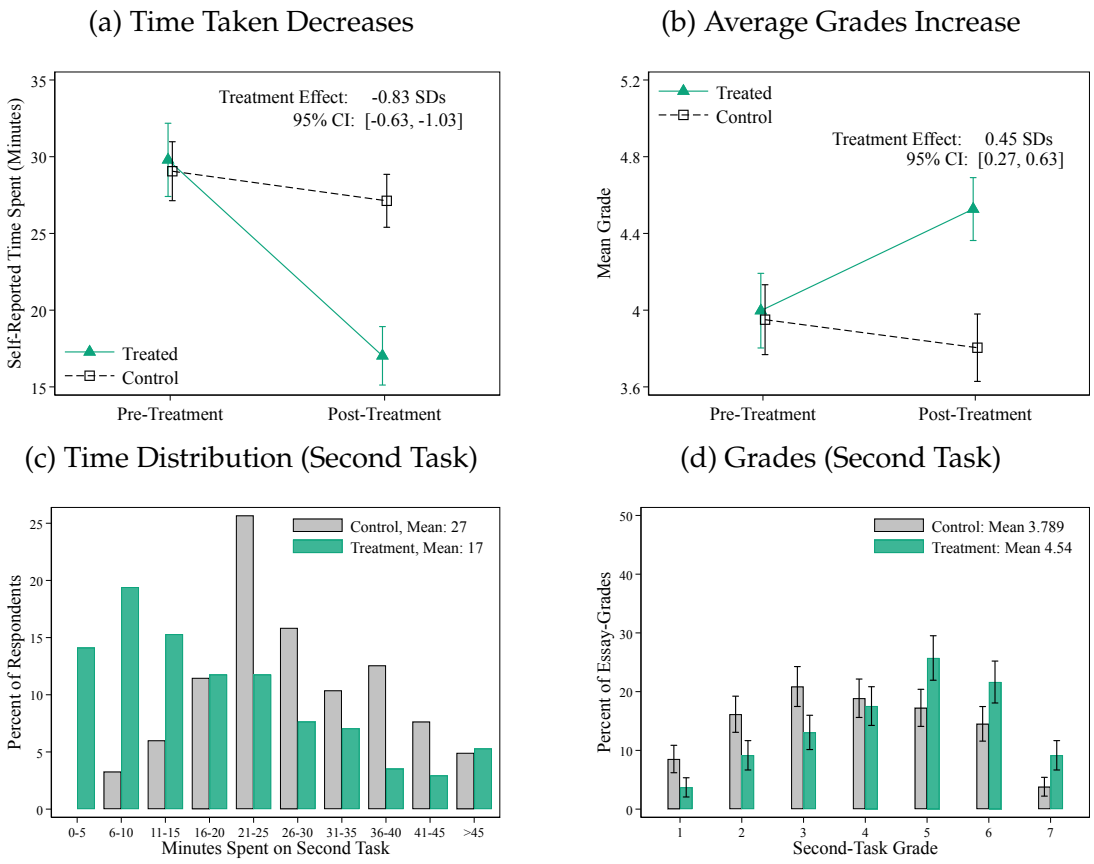
Figure 1 Panels (c) and (d) show that these effects are not limited to specific pockets of the time or grade distributions: the entire time distribution shifts to the left (faster work) and the entire grade distribution shifts to the right (higher quality). At the individual worker level, Figure 2 shows that treated workers who received a low grade on the first task experience both increases in grades and decreases in time spent, while workers who received a high grade maintain their grade level while substantially reducing their time spent.

These results are virtually identical across our two main incentive schemes, covering 80% of respondents: a "linear" scheme in which respondents are paid $1 for each point they receive

[2]Our survey was mostly active only after 5pm EST, when ChatGPT was almost always up, ensuring a very high successful signup rate. About 8% of treated respondents nevertheless face idiosyncratic technical difficulties that prevent them from signing up.

on each submission (each of which is graded on a 1-7 point scale), and a "convex" scheme in which respondents are additionally paid $3 for earning a grade of 6 or 7, giving them an extra incentive to produce high-quality output.

Figure 1: Treatment Effects on Productivity



Note: Panels (a) and (b) plot means (and 95% confidence intervals for those means) of self-reported time taken and average grades in the first and second task, separately in the treatment and control groups. The results look very similar for the objective measure of time active; see Supplementary Materials. The panels also display the coefficient on the treatment dummy from a person-evaluator-level OLS regression of the outcome variable (the within-person difference in the outcome between task 2 and task), on a treatment dummy, evaluator fixed effects, incentive arm fixed effects, and occupation*task-order fixed effects, clustering at the worker level. The treatment effect coefficient and standard error are normalized by dividing by the pre-treatment standard deviation of the outcome in the relevant incentive group(s). Panels (c)-(d) display raw graphs of the outcome distribution in the treatment versus control group on the second task.

Two supplementary interventions allow us to probe further. In one arm involving 20% of participants, we require participants in both the treatment and control group to spend exactly 15 minutes on each task. This holds effort fixed across the treatment and control groups, allowing us to interpret any difference in grades as a pure effect of ChatGPT access on productive capacity. In this arm, the treatment increases grades by a similar 0.39 standard deviations ($p = 0.13$), albeit imprecisely estimated and with a slight imbalance in pre-treatment

outcomes (see Online Appendix Figure A.7).

In another arm involving 30% of the treatment group, after completing the second task, respondents are shown their first-task output and given the opportunity to edit or replace it using ChatGPT if they wish. 23% choose to replace their response with ChatGPT's output and 25% use ChatGPT to edit their original response, suggesting that participants view ChatGPT as a way to improve output quality in addition to a convenient way to save time.

2.3 Productivity Inequality

The control group exhibits persistent productivity inequality: participants who score well on the first task also tend to score well on the second task. As Figure 2 Panel (a) shows, there is a correlation of 0.49 between a control participant's average grade on the first task and their average grade on the second task.

In the treatment group, initial inequalities are half-erased by the treatment: the correlation between first-task and second-task grades is only 0.25 (p-value on difference in slopes = 0.004). This reduction in inequality is driven by the fact that participants who scored lower on the first round benefit more from ChatGPT access, as the figure shows: the gap between the treatment and control lines is much larger at the left-hand end of the x-axis.

2.4 Human-Machine Complementarity

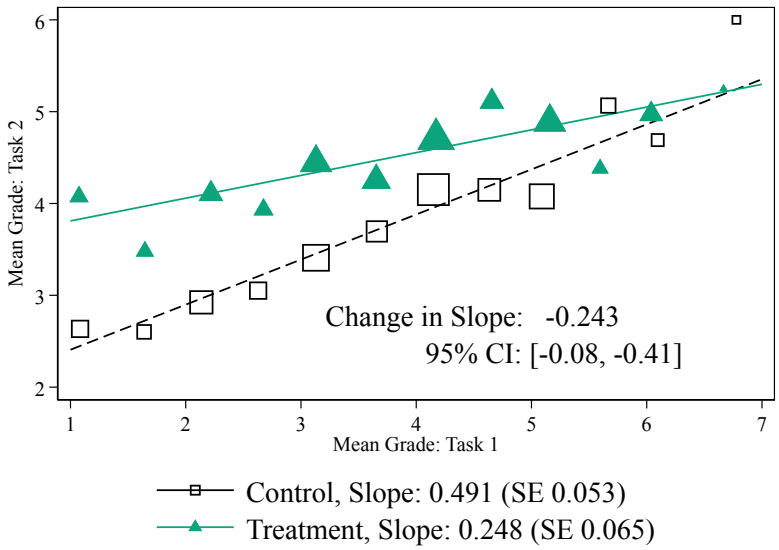
ChatGPT could increase workers' productivity in two ways. On the one hand, it could substitute for worker effort by quickly producing output of satisfactory quality that workers directly submit, letting them reduce the time they spend on the task. On the other hand, it could complement workers' skills: humans and ChatGPT working together could produce more than the sum of their parts, for example if ChatGPT aids with the brainstorming process, or quickly produces a rough draft and humans then edit and improve on the draft. In our experiment, evidence for the complementarity story could come in two forms: (a) we could observe treatment-group participants choosing to expend significant time editing ChatGPT's output or repeatedly prompting ChatGPT in anticipation of earning higher grades, and (b) we could observe that treatment participants' essays receive higher grades than ChatGPT's raw output, suggesting that human input adds value.

We observe neither of these pieces of evidence, suggesting that ChatGPT is increasing productivity primarily by substituting for worker effort. 68% of treated participants report submitting ChatGPT's initial output without editing it, and on average treated participants are active on the task for only 3 minutes after we first observe them pasting in a large quantity of text (presumably from ChatGPT). There is also no correlation between how long a participant is active after pasting in the ChatGPT text and the grade they ultimately receive, and treated

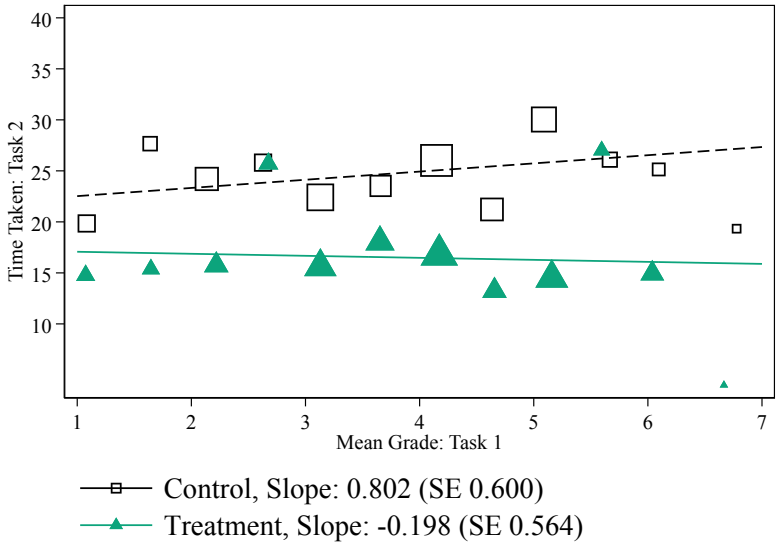
respondents do not receive higher average grades than raw ChatGPT output that we give to evaluators to grade, meaning we find no evidence that human editing is improving the ChatGPT output. This is true even when participants are given strong pecuniary incentives to do so, in the convex incentives group.

Figure 2: Effects on Grades and Time Across the Initial Grade Distribution

(a) Grade Inequality Decreases



(b) Time Taken Decreases Across Grade Distribution



Note: this figure display scatterplots, binning responses in equal intervals, of respondents' task-2 grade (Panel (a)) and task-2 time spent (Panel (b)) on their task-1 grade, separately by treatment and control group. Slopes are calculated through a worker-level regression.

2.5 Task Structure

As suggested by the preceding discussion, ChatGPT substantially changes the structure of writing tasks. Figure 3 Panel A shows that prior to the treatment, participants spend about 25% of their time brainstorming, 50% writing a rough draft, and 25% editing. Post-treatment, the share of time spent writing a rough draft falls by more than half and the share of time spent editing more than doubles.

2.6 Skill Demand

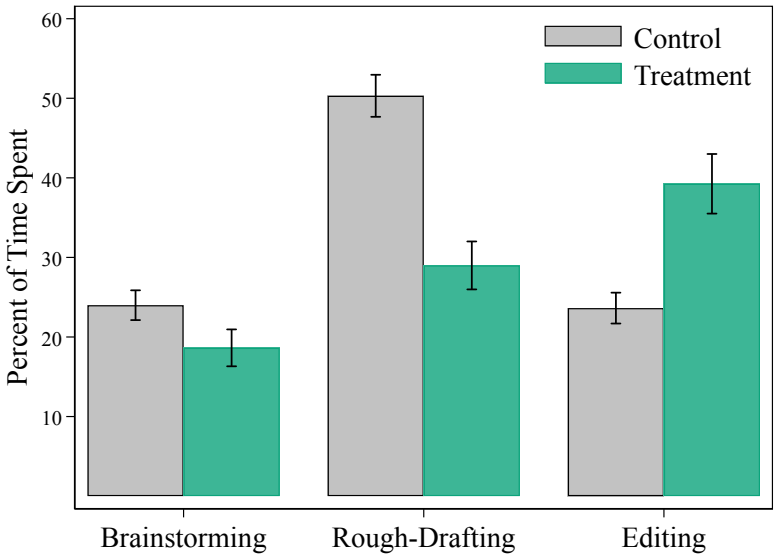
If ChatGPT is especially helpful to those with poor writing and communication skills relative to their other skills, it could have major labor market implications by expanding the available occupational choices and raising the earnings of individuals with strong idea-generation skills who struggle to effectively get those ideas onto paper.

We perform several tests of this hypothesis. We construct two measures of a person's relative writing skills. First, at the beginning of the experiment, we ask participants to rank from 1 to 3 their skills at communication (writing and speaking), problem solving, and creativity. Second, in addition to assigning overall grades, evaluators separately assess each piece of output based on writing quality, content quality, and originality; the gap between a person's first-task overall score and their writing score affords another measure.

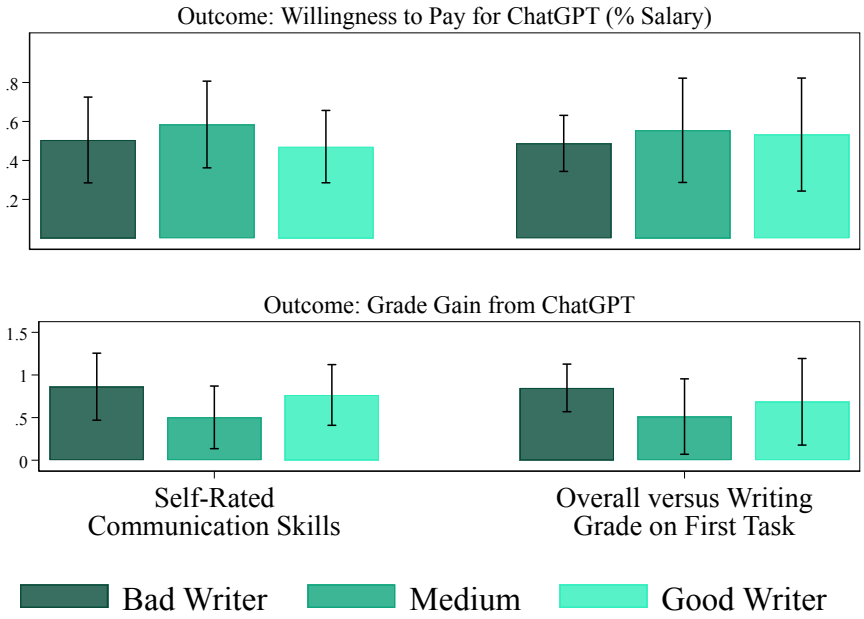
Similarly, we construct two measures of the individual-level benefits of ChatGPT. First, at the end of the experiment, we ask treatment-group participants how much they would be willing to pay on a monthly basis to access ChatGPT in their jobs. Second, we measure how much each treatment participant's grade increases from the first to the second task. We find no clear evidence for the aforementioned hypothesis. Figure 3 Panel B show that average willingness to pay for ChatGPT is flat across the terciles of both our measures of writing skill: respondents, regardless of their writing skills, are willing to pay about 0.5% of their monthly salary for a monthly subscription to ChatGPT. Grade gains from ChatGPT are also roughly flat across both measures of relative writing skills: people with comparatively poor writing skills do not experience unusually large grade gains.

Figure 3: Effects on Task Structure and Skill Demand

(a) Effects on Task Structure



(b) No Clear Heterogeneity in Benefits of ChatGPT by Relative Writing Skills



Note: Panel (A) shows self-reported time allocation on three separate task components, separately in the treatment and control group on the post-treatment task. Panel (B) shows that ChatGPT does not have greater benefits for treated respondents whose writing skills are poor relative to their other skills. Specifically, in the left-hand column, respondents are sorted by their self-assessed skill rankings: at the start of the survey, they rank their communication, problem-solving, and creativity skills. Those who rank communication 1st are defined as "Good" communicators, similarly for those who rank it 2nd and 3rd. In the right column, participants are sorted according to their grades: evaluators give each essay a separate overall, writing quality, content quality, and originality score. We define people as "Good" communicators if their average writing score exceeds their overall score, (about 30% of respondents), "Medium" if the two are equal (50%), and "Bad" if overall exceeds writing (20%). In the top row, the outcome is how much the participant is willing to pay to access ChatGPT in their job (elicited hypothetically). In the bottom row, the outcome is the participant's grade gain between the first and second task (everywhere we restrict to treated participants). Flat slopes within each of the

2.7 Job Satisfaction and Self-Efficacy

Access to ChatGPT could affect job satisfaction. For example, it could make participants happier by automating tedious or annoying components of the task or allowing them to finish more quickly. Alternatively, it could make the experience less enjoyable by quickly automating the most fun parts of the task. It could similarly either boost self-efficacy by giving participants access to a complex and powerful tool that enhances their capabilities, or it could lower it by making participants feel superfluous. We measure job satisfaction with a question, after each task, about how much participants enjoyed the task, and self-efficacy with a question about how skilled/effective they felt while completing the task, both on 1-10 Likert scales.

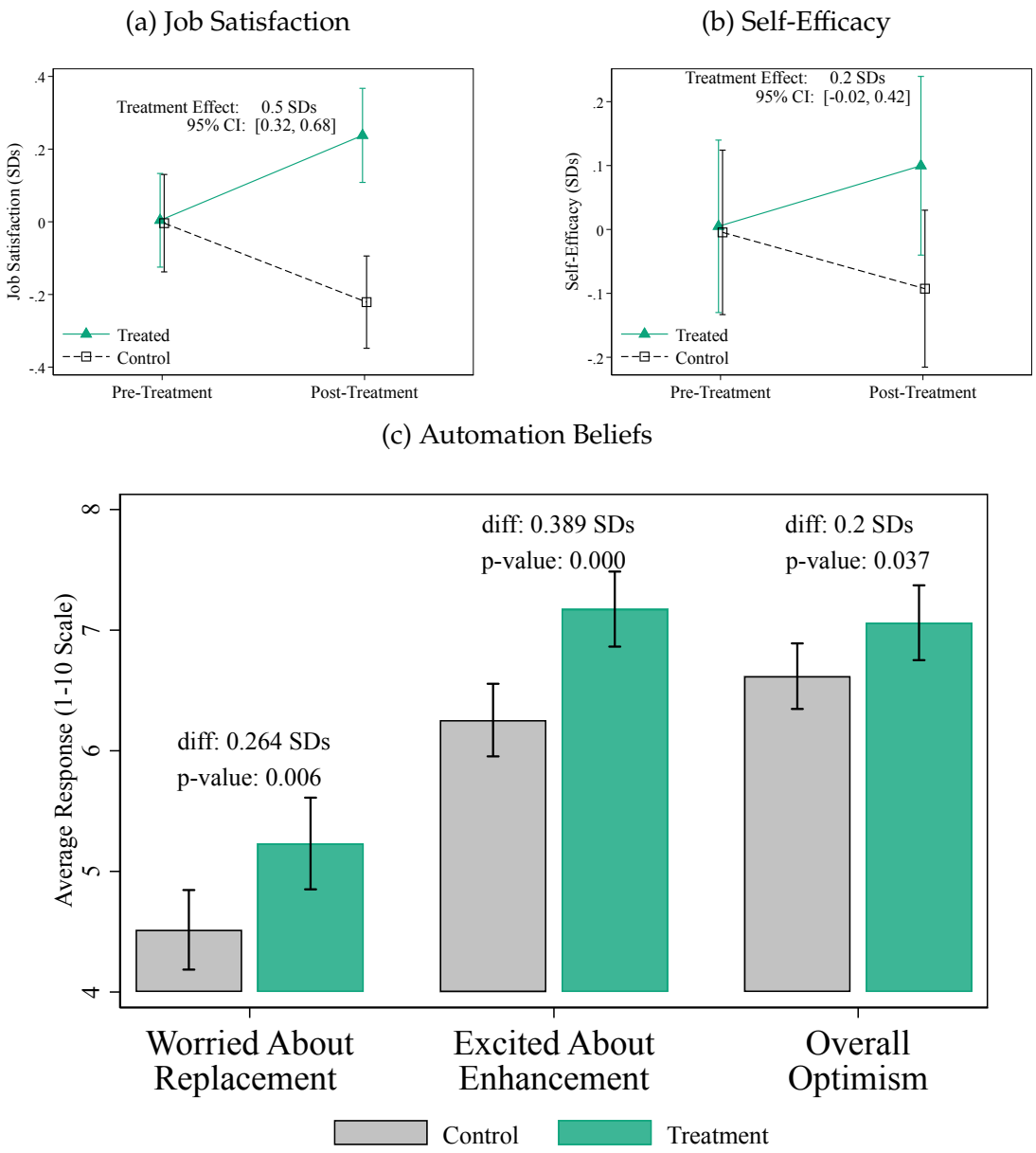
Figure 4 Panel (a) shows that ChatGPT substantially increases job satisfaction, by about 0.40 standard deviations (p = 0.000). Panel B shows that it mildly and imprecisely increases self-efficacy, by 0.20 standard deviations (p = 0.060), despite the fact that participants are mostly using it to substitute for their own effort. Qualitative feedback from participants (in an open-text box at the end of the survey) suggests that many enjoy discovering and working with this tool.

2.8 Beliefs About Automation

Many of our treated participants had never heard of (30%) or never used (70%) ChatGPT before participating in the experiment. Hence, most essentially encounter the technology for the first time and receive a crash course on its usefulness for writing tasks. How are their beliefs about future waves of automation affected by this encounter?

After respondents complete their second task, we elicit three beliefs, each on a 1-10 scale: how worried they are about workers in their occupation being replaced by AI; how optimistic they are that AI will make workers in their occupation more productive; and, overall, how optimistic or pessimistic they feel about future advances in AI. The effects of treatment on these outcomes are displayed in Figure 4 Panel (c); worry about automation increases by 0.26 standard deviations (p = 0.006), excitement by 0.39 (p = 0.000), and net optimism increases by about 0.20 (p = 0.037).

Figure 4: Effects on Subjective Outcomes



Note: Panels (a) and (b) show job satisfaction and self-efficacy (originally elicited on scales of 1-10, normalized to have mean 0 and standard deviation 1) pre- and post-treatment in the treatment and control group. Dots are means and error bars are 95% confidence intervals for means. The figures also print the coefficient on the treatment effect of a regression specified as in Figure 1. Panel (c) cross-sectionally compares beliefs about automation in the treatment and control group, all on 1-10 scales; the first question is "How worried are you about workers in your occupation being replaced by AI?" The second is "How optimistic are you that AI may make workers in your occupation more productive?" The third question is "How do you feel about the impacts of future advances in AI (1 = Very pessimistic, 10 = Very optimistic)".

2.9 Two-Week Followup Survey

One indication of the value of ChatGPT to participants is whether they continue to use it after the experiment. To track whether participants are subsequently using ChatGPT in their real jobs, we resurvey them two weeks after their completion of the initial survey. This followup is still in progress, with an 82% response rate among the 423 respondents who have been invited so far, and no evidence of differential response rates by treatment status.

33% of former treatment group participants have used ChatGPT in their job in the past week, relative to 18% of control group participants. Restricting to workers who had not previously used ChatGPT when they participated in our main experiment, 26% of treated and 9% of control workers are now using ChatGPT in their jobs (p-value on difference 0.048). Users give it an average usefulness score of 3.65/5.00, somewhat lower than in our main experiment, likely owing to the greater length and complexity of real-world tasks. The range of tasks they report using it for is broad: generating recommendation letters for employees, responding to customer service requests, brainstorming, search-engine requests, rough-drafting emails, and so on.

Respondents who are not using ChatGPT in their jobs mostly report that this is because the chatbot lacks context-specific knowledge that forms an important part of their writing. For example, they report that their writing is "very specifically tailored to [their] customers and involves real time information" or "unique [and] specific to [their] company products."

These comments point to an important (and inherent) limitation of our experiment: it involves relatively small, self-contained tasks that lack much context-specific knowledge beyond what we stipulate in the task prompts. However, our core result, that ChatGPT can increase productivity on many mid-level professional writing tasks, is supported by the fact that many respondents choose to use it in their real jobs. The fact that the treatment group is substantially more likely to use ChatGPT than the control group also suggests that the dissemination of ChatGPT into real professional activity is still in its very early stages, with many people not using it due to a lack of knowledge about or experience with the technology.

There is no difference between the former treatment and control in their overall satisfaction with their job in the followup survey. That said, many of our respondents only started using ChatGPT in their job in the past week or two, and it may take longer for access to ChatGPT to affect overall job satisfaction.

3 Discussion

College-educated professionals performing mid-level professional writing tasks experience substantial increases in productivity when given access to ChatGPT. The generative writing tool increases the output quality of low-ability workers while reducing their time spent, and it

allows high-ability workers to maintain their quality standards while becoming significantly faster. At the aggregate level, ChatGPT substantially compresses the productivity distribution, reducing inequality. It is also already being used by many workers in their real jobs. The experimental evidence suggests that ChatGPT largely substitutes for worker effort rather than complementing workers' skills, potentially causing a decrease in demand for workers, with adverse distributional effects as capital owners gain at the expense of workers.

The experiment has several important limitations worth enumerating. First, the tasks are relatively short, self-contained, and lack a dimension of context-specific knowledge, which may inflate our estimates of ChatGPT's usefulness. The results on job satisfaction and self-efficacy are similarly limited, reflecting enjoyment of a small task rather than feelings about a worker's whole job, as evidenced by the fact that there is no difference between the treatment and control groups in real job satisfaction after two weeks. Second, an experiment, by its nature, captures only direct, immediate effects of ChatGPT on the selected occupations. There will be many indirect, reinforcing, or counteracting "general-equilibrium" effects as labor markets and production systems adapt to the advent of technologies like ChatGPT. The effects of ChatGPT will also likely vary by occupation, task, and skill level.

Only time and future research will fully reveal how ChatGPT and its successors will affect labor markets. For now, the evidence we provide suggests that generative AI technologies will—and have already begun—to noticeably impact workers.

Bibliography

Acemoglu, Daron and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, 2020, *128* (6), 2188–2244.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb, "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction," *Journal of Economic Perspectives*, 2019, *33* (2), 31–50.

Autor, David, "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives*, 2015, *29* (3), 3–30.

_ **and David Dorn**, "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market," *American Economic Review*, 2013, *103* (5), 1553–1597.

Boustan, Leah Platt, Jiwon Choi, and David Clingingsmith, "Automation After the Assembly Line: Computerized Machine Tools, Employment and Productivity in the United States," *NBER Working Paper*, 2022.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," *Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, Skill, and Productivity: The Case of Taxi Drivers," *NBER Working Paper*, 2022.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Lee, David S, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Schwabe, Henrik and Fulvio Castellacci, "Automation, Workers' Skills and Job Satisfaction," *Plos one*, 2020, *15* (11), e0242929.

Table 1: Descriptive Statistics and Balance Tests

	Control			Treatment			
	N	Mean	SD	N	Mean	SD	Difference
Demographics							
Annual Salary ($)	215	71,808	47,648	197	76,267	51,271	4,458
Years of Tenure in Occupation	228	10.63	9.28	212	10.07	8.48	-0.564
Employed	220	0.90	0.29	207	0.96	0.20	0.052**
College Degree	229	1.00	0.07	215	1.00	0.07	-0.000
Occupation							
HR Professional	229	0.06	0.24	215	0.11	0.31	0.046*
Consultant	229	0.13	0.33	215	0.11	0.32	-0.015
Data Analyst	229	0.11	0.32	215	0.11	0.31	-0.007
Grant Writer	229	0.16	0.36	215	0.17	0.38	0.015
Manager	229	0.42	0.50	215	0.41	0.49	-0.014
Marketer	229	0.12	0.32	215	0.09	0.29	-0.025
First-Task Performance							
Time Spent Task 1 (minutes)	221	26.22	13.20	209	26.59	15.43	0.374
Grade Task 1 (1-7)	228	3.72	1.34	208	3.89	1.31	0.172
Job Satisfaction Task 1 (1-10)	228	6.32	2.54	212	6.34	2.36	0.019
Self-Efficacy Task 1 (1-10)	228	6.89	2.11	212	6.91	2.14	0.020

Note: This table plots means and standard deviations of descriptive characteristics for our sample, separately by treatment and control group. The right-hand side column shows the difference in means between treatment and control (levels of significance: * 10%, ** 5%, and *** 1%).