Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence

Authors: Shakked Noy[1*], Whitney Zhang[1]

Affiliations: [1]Department of Economics, Massachusetts Institute of Technology; Cambridge, MA, USA.

*Corresponding author. Email: snoy@mit.edu

Abstract: We examined 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 assigned occupation-specific, incentivized writing tasks to 453 college-educated professionals, and randomly exposed half of them to ChatGPT. Our results show that ChatGPT substantially raised productivity: average time taken decreased by 40% and output quality rose by 18%. Inequality between workers decreased, and concern and excitement about AI temporarily rose. Workers exposed to ChatGPT during the experiment were 2x as likely to report using it in their real job two weeks after the experiment, and 1.6x as likely two months after the experiment.

Main Text:

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 machine or computer, such as assembly-line manufacturing tasks or bookkeeping tasks (1,2). In contrast, creative, difficult-to-codify tasks (such as writing and image generation) largely avoided automation—a pattern scholars noted might change with the advent of deep learning, which now underpins generative AI systems.

The emergence of powerful generative AI technologies reintroduces a host of classic questions in a new context (3-5). Automation technologies—by definition—perform specific tasks in place of humans. But, more broadly, these technologies may either displace humans completely from certain occupations or augment existing human workers by increasing their productivity (6-9). If automation technologies (such as industrial robots) mostly displace human workers, they can increase unemployment. Moreover, their impacts on aggregate productivity may be small or nonexistent to the degree that they mainly serve to redistribute income previously earned by displaced workers to the capital owners supplying their robot replacements (10). If automation technologies (such as computers) augment existing workers, they can simultaneously benefit workers, capital owners, and consumers by raising wages, boosting productivity, and lowering prices (11-13).

A potent generative writing tool like ChatGPT could conceivably either displace or augment human labor. ChatGPT could 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. Alternatively, instead of displacing workers, 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 a rough draft. In this case, these services would become cheaper and demand could expand, resulting in higher employment and greater productivity for companies, cheaper products for consumers, and potentially higher wages for workers (14). Furthermore, inequalities between workers could either decrease if lower-ability workers are supported more by ChatGPT, or increase if higher-ability workers have the skills necessary to take advantage of the new technology.

Research Questions (RQ):

Which of these eventualities will generative AI systems bring about? The answer depends on a host of questions:

RQ1: How does access to generative AI systems affect workers’ productivity in existing tasks? Do workers choose to use these systems? Conditional on using these systems, how do workers interact with them and how do they affect productivity (15-18)?

RQ2: Do these systems differentially affect low- and high-ability workers?

RQ3: How do workers subjectively react to these technologies (19)?

Method:

This paper took the first step towards answering these questions (20). In a preregistered online experiment, we recruited 453 experienced, college-educated professionals on the survey platform Prolific and assigned each to complete two occupation-specific, incentivized writing tasks (21). The experiment took place from January 27th to February 21st, 2023, and involved GPT-3.5. The occupations we drew on were marketers, grant writers, consultants, data analysts, human resource professionals, and managers. The tasks, which included writing press releases, short reports, analysis plans, and delicate emails, comprised 20-to 30-minute assignments designed to resemble real tasks performed in these occupations; indeed, most of our participants reported completing similar tasks before and rated the assigned tasks as realistic representations of their everyday work (see Supplementary Materials).

Participants faced high-powered incentives, in the form of large bonus payments, to produce high-quality work: they received a base payment of $10 plus up to $14 in bonus payments for output quality, with the average overall rate of $17/hour substantially exceeding the Prolific standard of $12/hour. We cross-randomized the structure of bonus payments faced by participants to show robustness of our results to different incentive schemes (more details below). Output quality was assessed by blinded experienced professionals working in the same occupations. Evaluators were asked to treat the output as if encountered in a work setting and were incentivized to grade outputs carefully on a scale of 1-7 (22). Each piece of output was seen by three evaluators, with an average within-essay cross-evaluator correlation of 0.44 (23).

We randomly assigned 50% of participants to the treatment group and the remainder to the control group. The treatment group was instructed to register for ChatGPT between the first and second task, received guidance on using it, and were told they were permitted to use it on the second task if they found it useful. The control group was instead instructed to register for the LaTeX editor Overleaf, in an attempt to hold the time and hassle costs of signup constant between the two groups. The control group was not told they could use Overleaf on the second task and less than 5% subsequently reported using it.

In addition to output quality evaluations, we collected self-reported and objective measures of participants' time spent on the tasks and took snapshots of participants' outputs each minute while they performed the task, to construct objective measures of activity and detect ChatGPT usage (see Supplementary Materials).

A complete description of our experimental design, copies of relevant survey questionnaires, and additional figures validating our central measures and extending our main results are included in the Supplementary Materials. Descriptive statistics about the sample, as well as balance and selective attrition tests, are available in Table 1. The attrition rate was 6% in the control group and 11% in the treatment group. Balance tests indicate that across 13 pre-treatment characteristics, the treatment and control groups exhibited a small but significant difference only for only two characteristics: employment status and being an HR professional. Our partly within-person design, which controls for performance on the pre-treatment task, should eliminate any influence of selective attrition on our results; in the Supplementary Materials, we also report Lee bounds (24) 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.

Results

Takeup of ChatGPT

In the treatment group, 92% of treated participants successfully registered for ChatGPT, and 80% chose to use it on the second task (25). Users gave it an average self-assessed usefulness score of 4.4 out of 5.

Prior to treatment, 70% of our participants had heard of ChatGPT and 32% had used it before. Self-reported and objective measures indicate about 10-20% of the control group used ChatGPT on the tasks (see Supplementary Materials), meaning there was at least a 60-percentage point experimentally-induced gap in usage between our treatment and control groups on the second task. Our estimates reflect the effects of ChatGPT on the average productivity of the 60-70% of participants whose usage was determined by their treatment assignment, and constitute lower bounds on the effects of ChatGPT usage on productivity; in the Supplementary Materials, we report two-stage least squares results adjusting our estimates upwards for imperfect compliance.

Productivity

We first show results for our two productivity measures: time taken and evaluator grades (Figure 1). The experimental intervention shifted both outcomes dramatically. In the treatment group, time taken on the post-treatment task dropped by 11 minutes (0.75 standard deviations) relative to the control group, who took an average of 27 minutes (p<0.001). Average evaluator grades in the treatment group increased by 0.45 standard deviations (p<0.001), with similar increases for overall grades and specific grades for writing quality, content quality, and originality.

These effects are not limited to specific pockets of the time or grade distributions. As Figure 1 Panels C and D depict, the entire time distribution shifted to the left (faster work) and the entire grade distribution shifted to the right (higher quality). At the individual worker level, as depicted in Figure 2, treated workers who received a low grade on the first task experienced both 1-2 point increases in grades and 10-minute decreases in time spent, while workers who received a high grade maintained their grade level while also reducing their time spent by about 10 minutes.

These results are virtually identical across our two main incentive schemes, which covered 80% of respondents: a “linear” scheme in which respondents were paid $1 for each point they received on each submission (each of which was graded on a 1-7 point scale), and a “convex” scheme in which respondents were additionally paid $3 for earning a grade of 6 or 7. The results in Figure 1 are based on these two incentive schemes. The fact that treated participants reduced their time spent by a similar amount even when faced with strong incentives to produce high-quality output (under the convex scheme) demonstrates that the time-saving effects of ChatGPT are not specific to linear payment regimes and apply robustly across incentive structures.

In our third incentive arm involving 20% of participants, we required participants to spend exactly 15 minutes on each task, thereby holding effort fixed across the treatment and control groups and allowing us to interpret any difference in grades as a pure effect of ChatGPT access on productive capacity. In this arm, the treatment increased grades by a similar albeit not statistically significant 0.33 standard deviations (26).

In an additional intervention, after completing the second task, 30% of the treatment group were shown their first-task human-created output and given the opportunity to edit or replace it using ChatGPT. Of these participants, 19% chose to replace their response with ChatGPT’s output and another 17% used ChatGPT to edit their original response, suggesting that participants viewed ChatGPT as a means to improve output quality as well as save time.

Productivity Inequality

The control group exhibited persistent productivity inequality: participants who scored well on the first task also tended to score well on the second task. As Figure 2 Panel A shows, there was a correlation of 0.41 (*p*<0.001) between a control participant's grade on the first task and their grade on the second task, holding the evaluator constant.

In the treatment group, initial inequalities were more than half-erased by the treatment: the correlation between first-task and second-task grades was only 0.14 (*p*-value on difference in slopes<0.001). This reduction in inequality was driven by the fact that participants who scored lower on the first task benefited more from ChatGPT access, as the figure shows: the gap between treatment and control is much larger at the left-hand end of the x-axis.

Human-Machine Interactions

What kinds of human-machine interactions underlie the productivity results documented above? Did workers paste the task prompt into ChatGPT and immediately submit its output, minimizing their time spent and increasing their grades because ChatGPT's writing abilities exceeded theirs? Or did they treat ChatGPT as a helpful but imperfect tool, for example, using it to create a rough draft and then spending time editing and improving the draft, or using it to brainstorm or edit?

Our evidence supports the first possibility: almost everyone submitted lightly edited or unedited ChatGPT output, and we observed small time expenditures on editing and no resulting improvement in respondents' grades. In the treatment group, 33% of participants reported submitting ChatGPT's initial output without editing it, and 53% reported editing before submitting. However, those who reported editing were active on the task for only 3.3 minutes on average after we first observed them pasting in a large quantity of text (presumably from ChatGPT), with a majority active for 0-2 minutes (27). Qualitative examination suggests most of this editing was superficial, such as changing a placeholder or rearranging a sentence. Evaluator grades also suggest this editing was ineffectual: there was no correlation between how long a participant was active after pasting in the ChatGPT text and the grade they ultimately received, and treated respondents who used ChatGPT did not receive higher average grades than raw ChatGPT output we gave to evaluators to grade (see Supplementary Materials).

It is not obvious whether these dynamics should be interpreted as evidence that ChatGPT will displace human workers or evidence that it will augment them. On the one hand, ChatGPT directly substituted for participants' effort with little need for human input; on the other hand, it enabled participants to complete tasks much faster. We reflect on this further in the Discussion.

Subjective Outcomes: Job Satisfaction, Self-Efficacy, and Beliefs About Automation

Many of our treated participants had never heard of (30%) or never used (68%) ChatGPT before participating in the experiment. We used a battery of questions to assess their subjective reactions to encountering the technology. As depicted in Figure 3, participants enjoyed the tasks more by 0.5 standard deviations when given access to ChatGPT (*p*<0.001). Treated participants' concern for (*p*<0.01) and excitement about (*p*<0.001) future effects of AI on their occupations rose, and their overall optimism increased by 0.2 standard deviations (p<0.05). These effects disappeared in the two-week and two-month follow-up surveys, indicating they are best interpreted as short-run phenomena reflecting respondents' first experiences with the technology (28).

Two-Week and Two-Month Followup Surveys

One powerful indication of the value of ChatGPT to participants is whether they continued to use it after the experiment, in their actual jobs. To track this, we resurveyed participants two weeks and two months after their completion of the initial survey, with response rates of 92% and 83% respectively, and no treatment-control imbalance in response rates.

In the two-week follow-up, 34% of former treatment group participants reported using ChatGPT in their job in the past week, compared to 18% of control group participants (*p*-value on difference <0.001). Strikingly, this large gap in usage fully persisted into the two-month follow-up, where 42% of treatment and 27% of control respondents reported using ChatGPT in their job in the past week (*p*<0.01). The persistence of this gap suggests that the dissemination of ChatGPT into real professional activity is still in very early stages, with usage held back by lack of knowledge about or experience with the technology.

In the two-week follow-up, ChatGPT users gave the technology an average usefulness score of 3.66/5.00, somewhat lower than in our main experiment, likely owing to the greater length and complexity of real-world tasks. They reported using it for a broad range of tasks such as generating recommendation letters for employees, responding to customer service requests, brainstorming, rough-drafting emails, and editing.

Nonusers were divided into three roughly equal-sized groups, reporting either that: (a) ChatGPT was not useful in their job, (b) they did not know about it or did not have an account, or (c) it was not allowed in their workplace or usually unavailable during the day. The one-third of nonusers who claimed it was not useful in their job mostly said that this was because the chatbot lacks context-specific knowledge that forms an important part of their writing. For example, they reported that their writing was "very specifically tailored to [their] customers and involves real time information" or "unique [and] specific to [their] company products."

Discussion

College-educated professionals performing mid-level professional writing tasks substantially increased their productivity when given access to ChatGPT. The generative writing tool increased the output quality of low-ability workers and reduced time spent on tasks for workers of all ability levels. At the aggregate level, ChatGPT reduced inequality. It is already being used by many workers in their real jobs.

These results are consistent with other studies showing productivity-enhancing and equalizing effects of recent AI technologies (8, 15, 16, 18). Relative to these studies, we analyzed productivity effects across several occupations and tasks, examined how workers use ChatGPT, measured subjective reactions to the technology, and documented persistent effects of our treatment on ChatGPT usage in real jobs.

Limitations

The experiment had several important limitations. We examined a limited range of occupations and tasks, in which ChatGPT may be unusually useful. The tasks demanded clear, persuasive, relatively generic writing, which are arguably ChatGPT's central strengths. They did not require context-specific knowledge or precise factual accuracy. The version of ChatGPT used in this experiment cannot, by its nature, access or supply context-specific knowledge, and is not a reliable source of precise factual information.

The tasks could also be described through short, self-contained prompts, making use of ChatGPT easy, while many real-world tasks involve vaguer objectives and instructions, requiring workers to exercise initiative in determining what to do. Finally, participants in our tasks faced direct incentives in the form of bonus payments scaling with output quality, which encouraged them to maximize generic output quality and minimize time spent. White-collar workers are instead typically incentivized through longer-run promotion and firing incentives, which might instead encourage conspicuous exertion of effort or the development of a consistent personal style, both of which make ChatGPT less useful.

The tasks and incentive schemes were chosen to meet the constraints of the experimental design. We required short tasks that could be explicitly described for and performed by a range of anonymous workers online, and we needed to incentivize serious effort. Meanwhile, our judgement was that building factual-accuracy requirements into the tasks would either result in tasks that felt artificial and unnatural (for example, requiring participants to Google and report one or two specific facts), or overwhelm our budget for evaluators (for example, giving participants an open-ended research task and exhaustively fact-checking their assertions).

The aforementioned factors limit but do not eliminate the generalizability of our results. In real-world tasks, the need to fact-check ChatGPT's output will reduce its time-saving benefits, but the speed and writing quality increases observed in our experiment are sufficiently large that we suspect ChatGPT will still often be useful. Moreover, newer versions of ChatGPT are more consistently factually accurate and some versions can access the internet to fact-check themselves. We speculate that in more open-ended real-world tasks, workers may find iterative rounds of prompting and discussion with ChatGPT useful even if they cannot immediately prompt out a final product. In these contexts, ChatGPT and human workers may be more strongly complementary than in our experiment. The importance of context-specific knowledge will also limit ChatGPT's utility, but there are plausible workarounds: ChatGPT can be instructed to incorporate lists of context-specific factors, and organizations may be able to build customized ChatGPT-like models. Our follow-up surveys show that many workers do find ChatGPT useful in their real jobs.

Overall, we speculate that, relative to our experimental findings, the direct productivity effects of ChatGPT in the real economy will be somewhat lower and the technology will be more strongly complementary to human workers. To what extent remains an open question.

Implications

Our experiment captured only direct, immediate effects of ChatGPT on worker productivity. We could not examine the complex labor market dynamics that will arise as firms and workers adapt to ChatGPT. Several factors will mediate how the direct productivity impacts of ChatGPT affect wages and employment in exposed occupations.

The first is the degree to which demand for goods produced by ChatGPT could expand as ChatGPT-fueled productivity increases make those goods much cheaper. For example, demand for programming services could plausibly expand massively if the price of those services fell. Aggregate programming employment might consequently increase. It is less clear whether demand for advertising or communication could expand as much, potentially entailing a reduction of employment in those sectors as fewer workers aπre needed to meet the same static demand. As an additional complication, ChatGPT might directly affect the composition of demand. For example, prior to ChatGPT, a piece of writing signaled that a company had invested at least some human labor, thought, and judgement into a message, which consumers might have appreciated; with this no longer being the case, demand for these messages could decrease (29).

The second factor is the nature and scarcity of the human skills best complemented by ChatGPT. Consider use of ChatGPT to produce advertising content. Is this best accomplished by one senior advertising manager directly providing high-level guidance to ChatGPT, or by ten junior advertisers carefully designing prompts and editing ChatGPT's output? The answer will determine the structure of employment in the advertising sector. Similarly, suppose ChatGPT is highly complementary to human labor in programming tasks. If ChatGPT's human copilot needs to be an expert programmer capable of directly proofreading its output, this could raise programmers' wages by boosting their productivity while their expertise remains scarce. If, by contrast, the complementary human role requires only basic programming knowledge and mainly involves checking output and refining natural-language prompts, the pool of potential programmers would vastly increase, and wages could fall even as productivity rises. More generally, tools like ChatGPT could make expertise more accessible by facilitating learning (30).

Finally, the diffusion and effects of ChatGPT will also depend on organizational considerations that our experiment, treating isolated individual workers, does not speak to. ChatGPT might interact with traditional promotion and hiring systems based partly on conspicuous exertion of effort. Large language models may be used to monitor or evaluate workers and avoid paying higher wages (31). Organizational and societal norms around the acceptability of using tools like ChatGPT may take time to cohere and may significantly affect adoption of the technology (32-35).

Overall, the arrival of ChatGPT ushers in an era of vast uncertainty about the economic and labor market effects of AI technologies (36-38). Our experiment took the first step towards answering the many questions that have arisen.

References and Notes

1. D. H. Autor, Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives*. **29**, 3–30 (2015).
2. D. H. Autor, D. Dorn, The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*. **103**, 1553–1597 (2013).
3. T. Eloundou, S. Manning, P. Mishkin, D. Rock, GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models (2023), , doi:10.48550/arXiv.2303.10130.
4. E. W. Felten, M. Raj, R. Seamans, How will Language Modelers like ChatGPT Affect Occupations and Industries? (2023), doi:10.2139/ssrn.4375268.

5. M. R. Frank, D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. πFeldman, M. Groh, J. Lobo, E. Moro, D. Wang, H. Youn, I. Rahwan, Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*. **116**, 6531–6539 (2019).
6. L. P. Boustan, J. Choi, D. Clingingsmith, Automation After the Assembly Line: Computerized Machine Tools, Employment and Productivity in the United States (2022), , doi:10.3386/w30400.
7. D. Acemoglu, P. Restrepo, Robots and Jobs: Evidence from US Labor Markets. *journal of political economy*, 57 (2020).
8. K. Kanazawa, D. Kawaguchi, H. Shigeoka, Y. Watanabe, AI, Skill, and Productivity: The Case of Taxi Drivers (2022), , doi:10.3386/w30612.
9. R. Arakawa, H. Yakura, M. Goto, "CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023; http://arxiv.org/abs/2302.05678), pp. 1–19.
10. D. Acemoglu, P. Restrepo, The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review*. **108**, 1488–1542 (2018).
11. A. Agrawal, J. S. Gans, A. Goldfarb, Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*. **33**, 31–50 (2019).
12. M. Hoffman, L. B. Kahn, D. Li, Discretion in Hiring* | The Quarterly Journal of Economics | Oxford Academic. *Quarterly Journal of Economics*. **133**, 765–800 (2018).
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*. **133**, 237–293 (2018).
14. Productivity gains will translate into higher wages for workers if worker bargaining power or competition between employers for workers is sufficiently high to force employers to share part of the productivity gains with workers, and if the influx of workers into affected occupations is not large enough to completely offset these wage gains (see Discussion section).
15. E. Brynjolfsson, D. Li, L. R. Raymond, Generative AI at Work (2023), , doi:10.3386/w31161.
16. S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer, The Impact of AI on Developer Productivity: Evidence from GitHub Copilot (2023), , doi:10.48550/arXiv.2302.06590.
17. A. Calderwood, V. Qiu, K. Ilonka Gero, L. B. Chilton, "How Novelists Use Generative Language Models: An Exploratory User Study" in (ACM, Italy, 2018).
18. A. Campero, M. Vaccaro, J. Song, H. Wen, A. Almaatouq, T. W. Malone, A Test for Evaluating Performance in Human-Computer Systems (2022), (available at http://arxiv.org/abs/2206.12390).
19. H. Schwabe, F. Castellacci, Automation, workers' skills and job satisfaction. *PLOS ONE*. **15**, e0242929 (2020).
20. A nascent literature has studied applications of machine learning to *predictive* tasks (consisting of yes/no diagnoses) (11).
21. E. Peer, D. Rothschild, A. Gordon, Z. Evernden, E. Damer, Data quality of platforms and panels for online behavioral research. *Behav Res*. **54**, 1643–1662 (2022).
22. In each grading session, evaluators graded up to 14 responses. Evaluators received a base payment of $16, plus up to $8 in bonus payments depending on the correlation of their grades with the grades of other evaluators seeing the same responses.

23. This is the average correlation, across every pair of evaluators who saw the same essay, between the grade given by the first evaluator and the grade given by the second evaluator.
24. D. S. Lee, Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*. **76**, 1071–1102 (2009).
25. The choice to use ChatGPT is uncorrelated with treated respondents' salary, tenure, or grade on the first task.
26. Due to the small sample size in this group, the confidence interval includes zero and the treatment and control group differ in terms of average pre-treatment grades; see Supplementary Materials.
27. See Supplementary Materials for a full description of this analysis.
28. The effects on beliefs about automation may also have dissipated because rising global awareness of the technology led to an equalization of familiarity between the treatment and control groups.
29. Y. Liu, A. Mittal, D. Yang, A. Bruckman, "Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing" in *CHI Conference on Human Factors in Computing Systems* (ACM, New Orleans LA USA, 2022; https://dl.acm.org/doi/10.1145/3491102.3517731), pp. 1–13.
30. J. Qadir, Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education (2022), , doi:10.36227/techrxiv.21789434.v1.
31. D. Acemoglu, A. F. Newman, The labor market and corporate structure. *European Economic Review*. **46**, 1733–1756 (2002).
32. J. Ayling, A. Chapman, Putting AI ethics to work: are the tools fit for purpose? *AI Ethics*. **2**, 405–429 (2022).
33. J. Hohenstein, D. DiFranzo, R. F. Kizilcec, Z. Aghajari, H. Mieczkowski, K. Levy, M. Naaman, J. Hancock, M. Jung, "Artificial intelligence in communication impacts language and social relationships" (2021), , doi:10.48550/arXiv.2102.05756.
34. M. (Mia) Suh, E. Youngblom, M. Terry, C. J. Cai, "AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition" in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (ACM, Yokohama Japan, 2021; https://dl.acm.org/doi/10.1145/3411764.3445219), pp. 1–11.
35. H. Zohny, J. McMillan, M. King, Ethics of generative AI. *J Med Ethics*. **49**, 79–80 (2023).
36. D. Autor, The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty (2022), , doi:10.3386/w30074.
37. N. Kshetri, Artificial Intelligence in Developing Countries. *IT Professional*. **22**, 63–68 (2020).
38. A. Agrawal, J. Gans, A. Goldfarb, Eds., *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago London, First Edition., 2019).

Acknowledgments: We thank the editor, three anonymous referees, Daron Acemoglu, Nikhil Agarwal, David Autor, Lucas Barros, Talia Benheim, Amy Finkelstein, John Horton, Simon Jaeger, 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, Vod Vilfort, Shine Wu, and participants at the MIT Labor Lunch for helpful comments and conversations.

Funding:

Emergent Ventures grant, Mercatus Center, George Mason University (SN)

George and Obie Shultz Fund grant, MIT Economics Department (SN)

National Science Foundation Graduate Research Fellowship Grant 745302 (WZ)

Competing interests: Authors declare they have no competing interests.

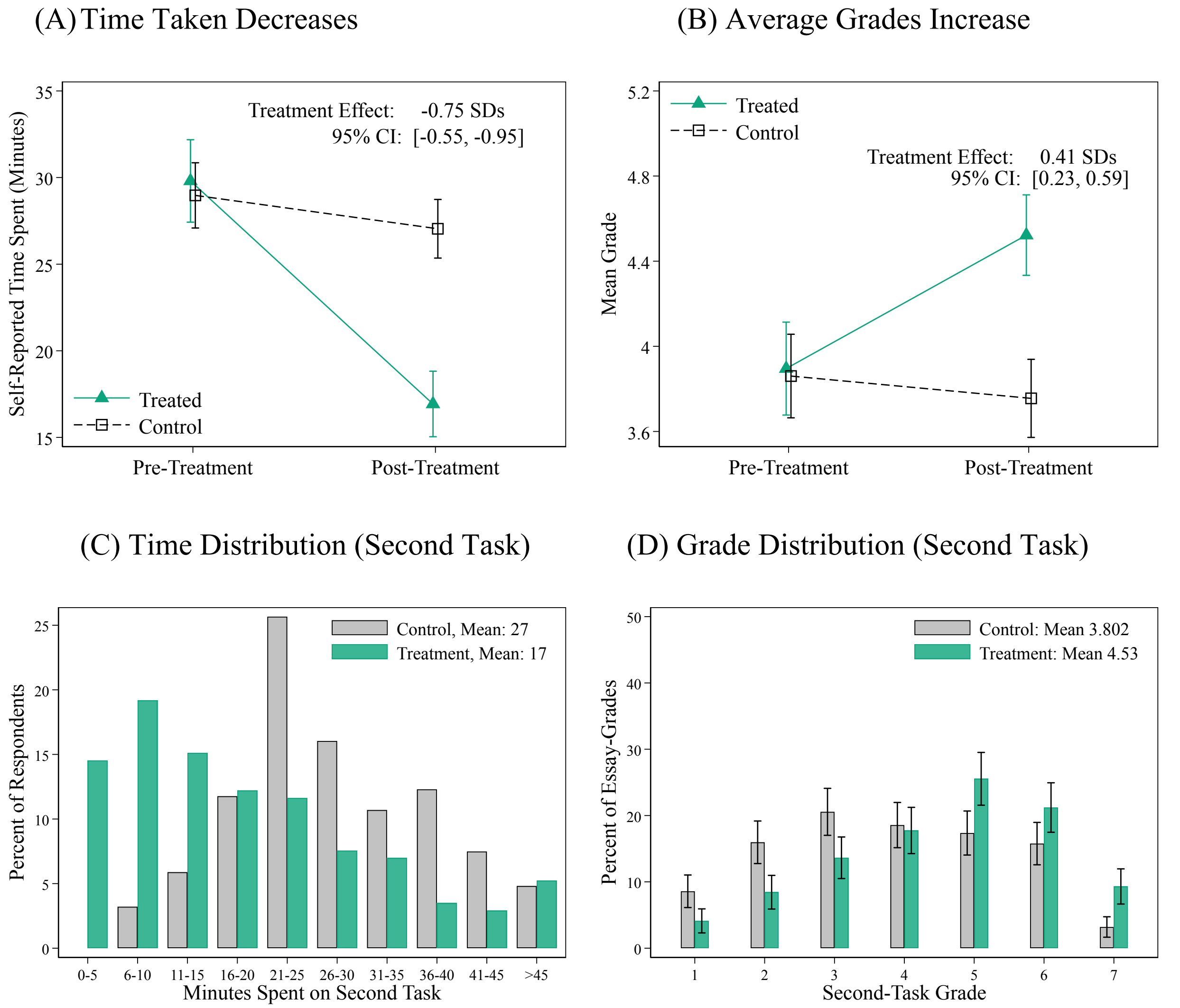
Data and materials availability: This experiment was preregistered at the AEA RCT Registry, https://www.socialscienceregistry.org/trials/10882

DOI https://doi.org/10.1257/rct.10882-2.1

All data and code used in the analysis are available on OSF at https://osf.io/xd7qw/?view_only=b8dac58dd6f44b979bf81069022fa392
DOI 10.17605/OSF.IO/XD7QW

Ethics: this experiment was approved by the MIT Committee on the Use of Humans as Experimental Subjects, Protocol Number 2212000849

Fig. 1. Treatment Effects on Productivity.



(A) Time Taken Decreases

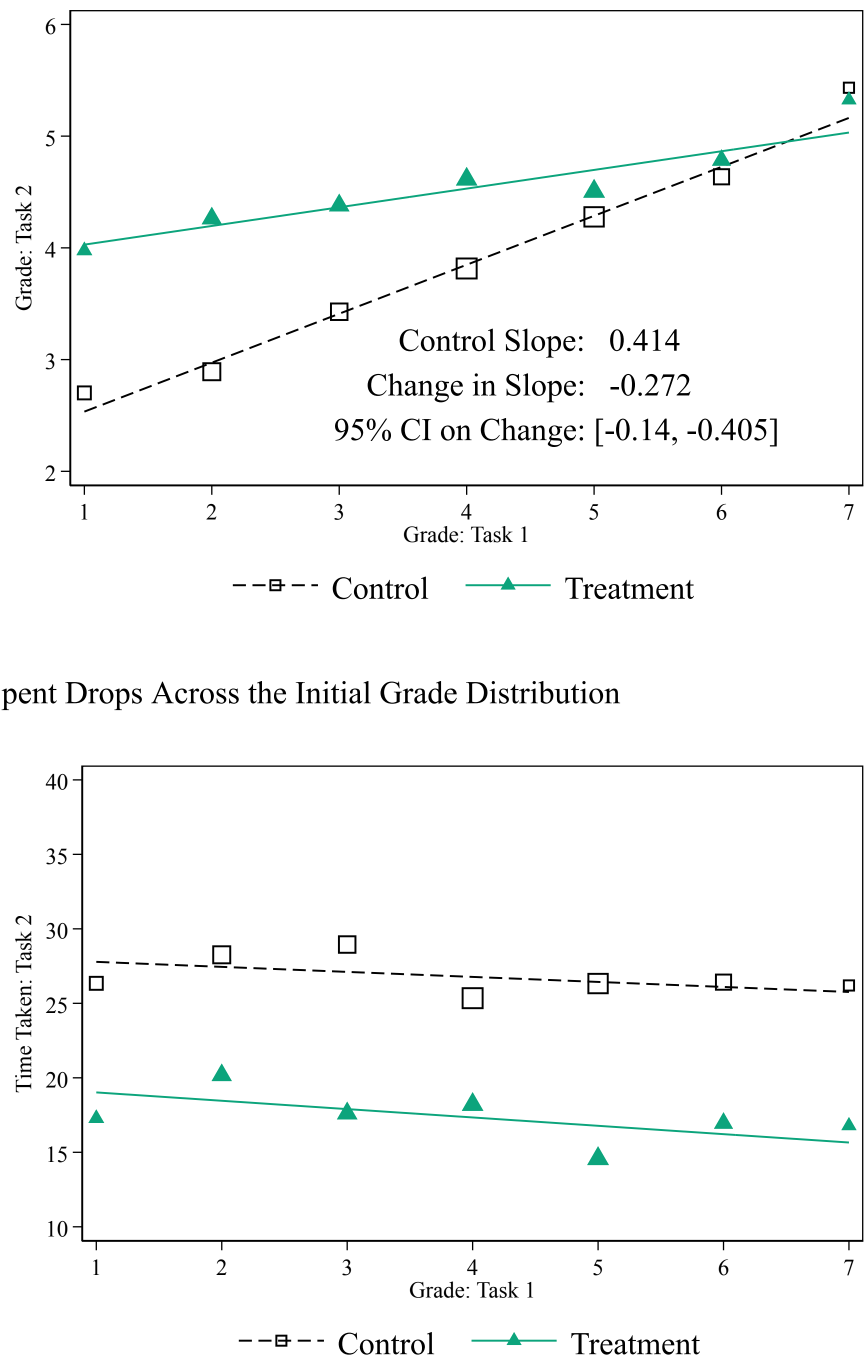
(B) Average Grades Increase

(C) Time Distribution (Second Task)

(D) Grade Distribution (Second Task)

*Note: All panels in this figure restrict to the linear and convex incentive groups. Panels (A) and (B) plots 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 treatment effect coefficients and 95% confidence intervals, rescaled to be in terms of pre-treatment standard deviations of the outcome variable. The coefficients are estimated from regressions of the within-participant change in outcome from pre-to-post treatment on a treatment dummy, occupation*task-order fixed effects, and incentive arm fixed effects. In Panel (A) this is at the participant level and standard errors are heteroskedasticity-robust. In Panel (B) this is at the participant-evaluator level, the regression also includes grader fixed effects, and standard errors are clustered at the person level. Panels (C)-(D) display raw* *graphs of the outcome distribution in the treatment versus control group on the second task. Panel (C) is at the individual level and Panel (D) is at the participant-evaluator level.*

Fig. 2. Effects on Grades and Time Across the Initial Grade Distribution.

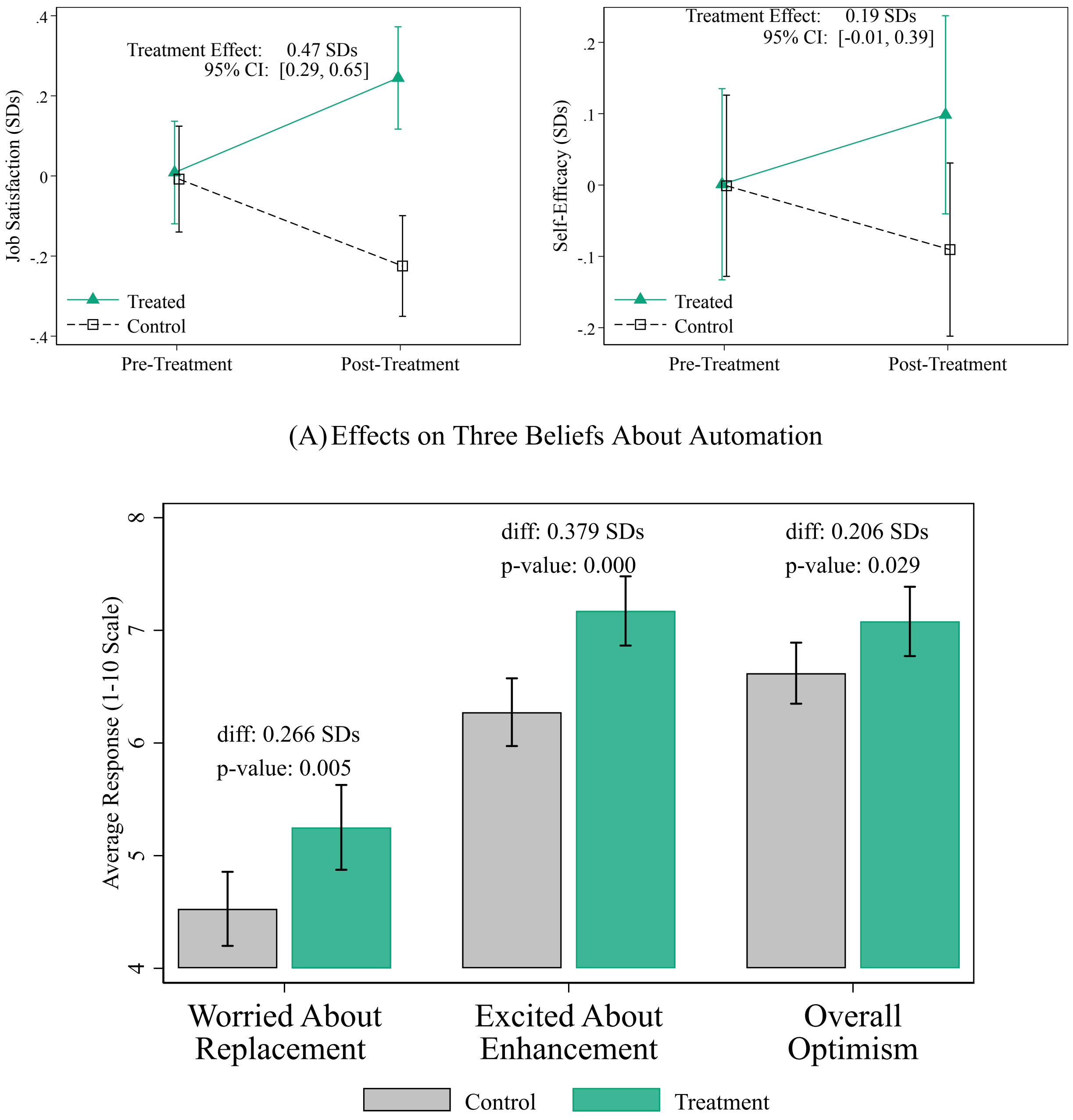


(A) Grade Inequality Decreases

(B) Time Spent Drops Across the Initial Grade Distribution

Note: these panels bin together participant-evaluator observations according to the task-1 grade given to this participant by this evaluator. Within each bin, the panels plot the average task-2 *grade (Panel A) or task-2 time taken (Panel B) for the observations in the bin, separately by treatment versus control. The panels also print the control-group slope, control-treatment difference in slopes, and a 95% CI for the difference; these latter results are calculated from a participant-evaluator level regression of the outcome variable on the task-1 grade, treatment status, treatment*task-1 grade, and grader fixed effects, clustering standard errors at the participant level. The control-group slope is the coefficient on the task-1 grade and the difference in slopes is the coefficient on the treatment*task-1 grade interaction. Note this* *difference in slopes will not match up exactly with the difference between the raw slopes plotted in the graph, since these raw slopes do not use grader fixed effects.*

Figure 3. Job Satisfaction, Self-Efficacy, and Beliefs about Automation.



(A) Job Satisfaction Increases

(B) Self-Efficacy Marginally Increases

(A) Effects on Three Beliefs About Automation

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 A. 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)".

Table 1. Descriptive Statistics.

Variable	N (Control)	Mean (Control)	N (Treatment)	Mean (Treatment)	Difference (*p<0.10, **p<0.05, ***p<0.01)
Annual Salary in Main Job ($)	234	67,764	213	71,938	4,173
Years of Tenure in Occupation	234	10.49	215	10.07	-0.43
Employed	226	91%	210	96%	5.0%**
Occupation: HR Professional	235	6%	218	11%	4.6%*
Occupation: Business Consultant	235	13%	218	11%	-1.3%
Occupation: Data Analyst	235	11%	218	11%	-0.0%
Occupation: Grant Writer	235	16%	218	17%	1.2%
Occupation: Manager	235	43%	218	41%	-1.7%
Occupation: Marketer	235	11%	218	9%	-2.3%
Time Spent (Task 1, Minutes)	227	26.10	212	26.58	0.47
Average Grade (Task 1)	233	3.63	211	3.77	0.15
Job Satisfaction (Task 1, 1-10 Scale)	234	6.30	215	6.34	0.04
Self-Efficacy (Task 1, 1-10 Scale)	234	6.89	215	6.90	0.01

Notes: this table presents descriptive statistics for our sample. We recode salary reports of >$500,000 to missing (affects 2 observations). "Employed" includes fulltime and parttime.