

The Cost of Generative AI Collaboration Strategies to the Psychological Experience of Work

Jennifer M. Logg*

McDonough School of Business, Georgetown University

Matthew R. Leitaio*

National Institutes of Health

Canan Ulu

McDonough School of Business, Georgetown University

*Denotes co-first authorship.

Working Paper: June 12, 2026.

Author Contributions: JL: Conceptualization, Methodology & Pre-Registration, Investigation, Writing - Original Draft, Writing - Review Editing, Supervision, Project Administration, Funding Acquisition; ML: Software, Formal Analysis, Investigation, Data Curation, Visualization; CU: Conceptualization, Methodology Pre-Registration, Investigation, Writing - Original Draft, Writing - Review Editing, Supervision, Project Administration, Funding Acquisition.

Author Note: The Authors wish to thank the AI, Analytics, and the Future of Work Initiative at Georgetown University's McDonough School of Business for their generous support in funding this research. Correspondence concerning this article should be addressed to Jennifer M. Logg, McDonough School of Business, Georgetown University, Washington, DC 20057. E-mail: jenn.logg@georgetown.edu.

Abstract. Generative AI (GenAI) is transforming the way people work by increasing productivity and changing workflows. Yet, little is known about the optimal way to use it in order to balance both productivity and the psychological experience of work (i.e., control, responsibility, and creativity). We examined different strategies, which we call *GenAI collaboration strategies*, and their associated productivity benefits and psychological costs. In two pre-registered experiments ($N = 1,299$), we examined two common uses of GenAI, using it to draft versus edit. First, using GenAI, relative to not using it, improved productivity but degraded control, responsibility, and creativity. Second, we identified the optimal GenAI collaboration strategy: using GenAI to edit. Productivity was similar between collaboration strategies, but using GenAI to draft further hurt people's psychological experience of work. Yet, the majority of people chose the suboptimal strategy, including a sample of teachers creating lesson plans for a class. GenAI is a powerful tool to increase productivity, but it is not without costs. Thus, it is not enough for organizations to provide employees with GenAI; how people use GenAI has implications for their psychology.

Key words: Generative Artificial Intelligence (GenAI), AI Use, Productivity, Psychological Experience of Work, Cost-Benefit, Experimental.

Introduction

The rapid advancement of generative artificial intelligence (GenAI) has the potential to change work and productivity in virtually all economic sectors. The promise of innovation through GenAI is in automating a broad variety of tasks, tasks that technology could not do a few years ago. In fact, GenAI and automation are expected to displace 92 million jobs and create 170 million new jobs by 2030 (World Economic Forum 2025). Employees are using GenAI tools that leverage Large Language Models (LLMs), such as ChatGPT, Claude, and Gemini, to help with writing and editing (creating reports and proposals), programming (generating code or debugging), design and content creation (producing visual design and branding), customer support, and

data analysis. Organizations are currently trying to materialize productivity gains from GenAI by creating higher quality deliverables at a faster rate.

Understanding if GenAI impacts the productivity of work is the first step towards leveraging its benefits. Extant research on the use of GenAI focuses on the quality (Dell’Acqua et al. 2026) and productivity benefits of GenAI assistance, relative to its absence (Noy and Zhang 2023, Qian and Wexler 2024, Brynjolfsson et al. 2025). However, in many cases, employees are no longer focused on *whether* or not to use GenAI but, *how* they will use GenAI. In fact, the concept of AI-Human “teaming” (AI-Human collaboration) has become so popular in organizations (Zao-Sanders 2025, Alfaro et al. 2026, Gonzalez et al. 2025, 2026) that many take the concept for granted.

GenAI Collaboration Strategies. Managers can benefit their organization by shifting their focus from *whether* to adopt GenAI to *how* they can help their employees identify their optimal collaboration strategies. Although there is some descriptive data about which tasks employees use GenAI (Chatterji et al. 2025, Zao-Sanders 2025), we know less about how they are using it within their workflow for those tasks (Cu and Hochman 2023, Den Houter 2024). For example, 44% of Chief Human Resources Officers do not know how their employees currently use algorithmic tools (Den Houter 2024, Brecheisen 2024).

We examined specific strategies, which we call *GenAI collaboration strategies*, using GenAI to draft initial work (GenAI Drafter) versus using GenAI to refine existing work (GenAI Editor). We examined writing tasks, which represent the most prevalent professional application of GenAI (40% of work-related usage, Zao-Sanders 2025, Chatterji et al. 2025). More specifically, we examined drafting and editing as they are the fundamental tasks of writing (Flower and Hayes 1981). While Chen and Chan (2024) compared two different GenAI modalities with a control condition (not using GenAI), they did not compare the modalities with each other. In the GenAI “ghostwriter” modality, GenAI drafted an ad copy based on instructions from the participant. In the GenAI “sounding board” modality, GenAI provided feedback on the participant’s draft but was not allowed to draft. We extend this work by identifying an optimal strategy and comparing both drafting and editing with a control (no GenAI) as well as with each other. Thus, we tested: Does it matter how people utilize GenAI? Are there costs associated with different strategies?

Costs of GenAI Collaboration. Are there tradeoffs to collaborating with GenAI, or are such collaborations purely beneficial? We take a cost-benefit perspective to the use of GenAI. While there is considerable discussion on the potential benefits of using GenAI in the workplace (Tamkin and McCrory 2025), we know little about the potential costs. Past work shows that the presence (versus absence) of GenAI improves productivity (Brynjolfsson et al. 2025, Noy and Zhang 2023, Qian and Wexler 2024). But as organizations face the allure of GenAI’s productivity benefits, they simultaneously face chronically low employee engagement, that has developed into a costly culture of “quiet quitting.” Employees have lost so much motivation that they disengage from their work, culminating in a broader “employee engagement slump.” This slump cost

the global economy \$10 trillion in productivity in 2025 and \$438 billion in 2024 (Gallup 2025, 2026). We cannot understand the full impact of GenAI on employees until we examine both the potential costs and benefits.

What are the psychological effects of GenAI on people? Might collaborating with GenAI degrade employees' psychological experience of work? If so, organizations focused solely on the benefits of productivity risk degrading the integrity of the entire work process. Such degradation is meaningful to organizations even as they automate more tasks and roles; GenAI-human collaboration will remain necessary for many jobs. First, human judgment is necessary to discern which data is most relevant to organizational goals and problems and to determine whether the output is appropriate and relevant for those goals and problems (Luca et al. 2016, Logg 2022). Second, although automation may replace some jobs, it will likely create new ones (World Economic Forum 2025). In these situations, the person using GenAI may be the person bearing the responsibility of executing the final task or leading their team to do so (i.e., the lawyer who defends the case, the teacher who teaches a class, or the judge who renders a verdict). If a teacher uses GenAI to produce lesson plans but does not feel responsible enough for the topics to implement the plans, this collaboration fails the teacher, the students, and the school. If a manager uses GenAI to create a plan for a new product launch, but feels little control over that process, they may fail in leading their team to deploy the plan. Ultimately, the company may fail to bring the product to market.

Moreover, while GenAI's performance is currently improving on tasks such as text generation and coding, there is still a significant gap in the quality of certain tasks completed by GenAI and human experts. (Mazeika et al. 2025) found that the best AI system successfully completed only 2.5% of the assignments given to people on work project platforms (creating dashboards for data visualization, 3D product videos, and games). In fact, despite the rise in automation since the introduction of GenAI, customer service job postings in the U.S. have increased since the summer of 2025 (Indeed 2026), as people are still needed to correct mistakes made by automation and to address escalated issues. This suggests that collaboration between individuals and GenAI is still necessary for high-quality output.

Optimal GenAI Collaboration: Psychological Experience of Work. If people are not engaged in their jobs, GenAI is less helpful for organizational success. Indeed, recent work suggests that the presence of GenAI (relative to the absence) decreases motivation (Wu et al. 2025). Beyond this work, we know very little about the costs associated with GenAI to the workflow and to the worker. Moreover, to our understanding, there is no work examining the costs associated with specific collaboration strategies. Examining the costs of these strategies allows research to identify an optimal strategy for using GenAI that takes into account both productivity and the psychological experience of work. Might certain collaboration strategies benefit workers more than others? Are certain strategies associated with greater costs?

Productivity is not the only goal of most employees, or even organizations (Danna and Griffin 1999, Grant et al. 2007). Even when productivity is the end goal for the organization, it is clear from decades

of research that the psychological experience of work influences employees' performance at their jobs. In fact, the Job Characteristics Model shows that feelings of responsibility for work are associated with motivation, satisfaction, and performance (Hackman and Oldham 1976, Humphrey et al. 2007). Similarly, research shows that creativity is associated with higher job performance, as rated by supervisors (Gong et al. 2009, Amabile and Pratt 2016). Thus, a complete cost-benefit analysis of GenAI must account for how collaboration strategies affect workers' psychological experience, not just their productivity. To capture different dimensions of the psychological experience of work, we drew from this literature and measured feelings of control, responsibility, and creativity.

How does collaborating with GenAI influence how people see their engagement with their work? Do certain GenAI collaboration strategies make people feel less responsible for their work, feel less control over their work, or feel less creative? Measuring both productivity and the psychological experience of work allowed us to test which strategy was optimal when taking both goals into account. For our main research questions, we adopted a cost-benefit perspective to examine the use of GenAI:

Research Question 1. Is there an optimal GenAI collaboration strategy?

Examining specific strategies allowed us to measure people's preferences for how they use GenAI. This work is the first to our knowledge to measure people's preferences for some of the most common uses of GenAI, to draft versus to edit. Previous work randomly assigned people to the presence versus absence of GenAI assistance (Brynjolfsson et al. 2025, Noy and Zhang 2023, Qian and Wexler 2024) or to a particular modality (Chen and Chan 2024). We extended this work in Experiment 2 by capturing choice, allowing us to evaluate whether people select the optimal GenAI collaboration strategy. Our second and third research questions were:

Research Question 2. Do people have a preference for how to use GenAI (GenAI collaboration strategy)?

Research Question 3. Do people select the optimal GenAI collaboration strategy?

Finally, examining specific strategies also allowed us to uncover the mechanisms that unfold within the collaboration. This extends extant work by testing what happens during these collaborations and why there is an optimal GenAI collaboration strategy. We measured GenAI and people's involvement in the collaboration through their editing patterns and time spent in each stage of the workflow. Our fourth research question was:

Research Question 4. What are the mechanics of how people collaborate with GenAI for specific GenAI collaboration strategies?

The Present Research: Balancing the Costs and Benefits of GenAI

Does the productivity benefit from injecting GenAI into employees' workflows come at a cost to their psychological experience of work? Are there differential costs and benefits associated with different collaboration strategies? Across two pre-registered experiments ($N = 1,299$), we examined how GenAI collaboration

strategies (using GenAI to draft versus edit) influenced people's productivity as well as their psychological experience of work.

To identify an optimal GenAI collaboration strategy (RQ1), we randomly assigned people to a strategy and measured both productivity and their psychological experience of work (Experiment 1). To test whether people selected the optimal strategy (RQ2 and RQ3), we randomly assigned people to 1) a GenAI condition in which they chose whether GenAI would draft or edit the deliverable, or 2) to a control condition (Experiment 2).

In order to understand the dynamics of the collaboration and to unpack how each strategy influenced productivity, we examined how much people and GenAI actively contributed to the collaboration (RQ4, Experiments 1 & 2). We compared the magnitude of edits made by people to drafts produced by GenAI with the magnitude of edits made by GenAI to drafts produced by people. Then, we compared the deliverables using text distance measures.

Across both experiments, we found that simply including GenAI in one's workflow is not enough; how people used GenAI mattered. In Experiment 1, participants wrote podcast proposals. Counter to our pre-registered predictions, the strategies similarly influenced productivity (and did not interact with expertise). We examined why productivity was similar between strategies by separating quality and time. Importantly, the strategies did have a differential effect on people's psychological experience of work. Using GenAI to draft degraded people's psychological experience of work. Specifically, participants who were randomly assigned to use GenAI to draft their work experienced less control, responsibility, and creativity. We next tested if these results held when people were able to select their GenAI collaboration strategy.

In Experiment 2, we tested people's GenAI collaboration strategy preferences and recruited a sample of teachers to create a lesson plan that they could use for their jobs. We selected this sample and task in order to increase the ecological validity of the task and to increase the likelihood that participants were motivated to optimize for quality as well as time. This helped rule out that the preference to use GenAI as a drafter was driven solely by the goal to save time.

We found that the majority of teachers selected to use GenAI to draft. We also found that using GenAI improved productivity over not using GenAI, without any productivity differences between collaboration strategies. Again, how people used GenAI influenced whether people optimized their work: using GenAI as a drafter degraded their psychological experience of work without any added productivity benefit. Thus, paradoxically, when given the choice, people selected the suboptimal GenAI collaboration strategy. Additionally, we examined perceptions of these strategies: people's frustration with them, expectations about using both the strategy and the lesson plan in the future, and reputational concerns with the strategy. Both experiments showed that using GenAI as a drafter degraded people's psychological experience of work without additionally benefiting productivity, compared with using GenAI as an editor. Organizations cannot

simply provide access to GenAI, otherwise they risk employees selecting a strategy that alienates them from their work.

Open Science Statement. We reported all of the pre-registered sample sizes, data exclusions, manipulations, and measures (Simmons et al. 2012). Pre-registrations including exclusions (Logg and Dorison 2021), materials, data, and supplemental materials are posted online at the Open Science Framework: https://osf.io/eaznb/overview?view_only=0d7132aa04764803b4b2c01fe2609ab6. For both experiments, we used ChatGPT Model 4.5. For our analyses, we used R [Version 2026.01.0+392] (R Core Team 2025), the Psych Package [Version 2.6.1] (Revelle 2026), and the BayesFactor package [Version 0.9.12-4.7] (Morey and Rouder 2024).

1. Experiment 1: Podcast Proposal

Is there an optimal GenAI collaboration strategy (RQ 1)? In Experiment 1, we tested whether different GenAI collaboration strategies (using GenAI as a drafter or editor) affected people’s productivity and psychological experience of work. This allowed us to test whether certain strategies boost productivity at a cost to the psychological experience of work (control, responsibility, and creativity). Additionally, we tested: What are the mechanics of specific GenAI collaboration strategies (RQ4)? To test what drives differences in quality and total time, we measured the length of deliverables and the amount of editing (the change in length from the initial to the final version) for each GenAI collaboration strategy. Additionally, we examined whether topic expertise influences the effects of GenAI collaboration strategies on productivity and the psychology of work. Although the topic expertise manipulation was successful, it did not affect our outcomes.

1.1. Method

Participants. The final sample included 601 participants (352 women; 237 men; 12 other; median age = 37) from Prolific. Participants were compensated with \$1.60 for the 8-minute survey. They had the opportunity to receive a bonus up to \$1.50, based on how third-party judges scored their deliverables. Each deliverable (both the initial and final draft) was scored by three judges ($N = 1202$; who each scored three deliverables). We used the average scores of the final draft, the average judge score, for bonus payments.

Design. We employed a 2 (Topic Expertise: High vs. Low) $\times$ 2 (GenAI Collaboration Strategy: GenAI as Drafter vs. GenAI as Editor) between-subjects design. After rating their knowledge of topics from a list, participants were randomly assigned to the topic they knew most about (high topic expertise) or least about (low topic expertise) and then randomly assigned to a GenAI collaboration strategy.

Procedure and Materials. Participants were asked to write a short pitch for a new podcast. We purposely chose a task that required persuasive communication and creativity in order to understand the costs and benefits to people when collaborating with GenAI in a more subjective domain. Participants first read a description of the writing task, “For this survey, please generate a short pitch (200-250 words) for a new

podcast. Podcasts cover many topics, from news and business to entertainment and comedy to science and technology. We want you to pitch a new podcast and describe the kinds of episodes you would include.” On the next page, they read that they had the opportunity to receive a bonus payment based on “how participants in a separate survey judge your pitch.” They read that these three judges would use four criteria (see Table 1) to score each pitch for a total possible score of eight and that the bonus depended on the average judge score.

Table 1 Evaluation criteria in Experiment 1

Criteria	Explanation
1	Persuasiveness of pitch
2	Creativity of pitch
3	Feasibility of pitch
4	Amount of interest your podcast would receive from listeners

Next, participants read a list of thirteen topics (e.g., ancient civilizations, architecture, pop culture) and ranked them in order from those they felt the most knowledgeable about to those they felt the least knowledgeable about. On the following page, participants read their assigned topic and read information about how they would interact with GenAI. Participants were randomly assigned to either the topic they reported knowing the most or the least about. They were also randomly assigned to either have GenAI draft the pitch before they edit it or draft the pitch themselves before GenAI edits it.

In both the GenAI editor and GenAI drafter conditions, participants were shown the prompt that was used to interact with GenAI and did not iterate with GenAI beyond this single interaction. Participants who were assigned to use GenAI to edit were asked to draft their podcast pitch. Their topic, draft, and a prompt to edit according to the judging criteria were sent to GenAI. Participants were shown GenAI’s edited plan and, without further modification, submitted that as their final lesson plan for scoring. Participants who were assigned to use GenAI to draft read GenAI’s draft. They could make as many or as few edits as they wanted in order to produce a final pitch, which was then submitted for scoring. Participants also estimated the average judge score that their podcast proposals would receive immediately after the initial and final pitches. Next, they answered questions about their psychological experience of work while writing their pitch, individual differences, and demographic information, detailed below.

Psychological Experience of Work. In order to understand if participants’ psychological experience of work was impacted by the GenAI collaboration strategy, we measured three key dimensions of the psychology of work: control, responsibility, and creativity. Participants answered the following: “How much **control** do you feel you had over the pitch?” on a scale from 0% (*No Control*) to 100% (*Complete Control*), “How much **responsibility** do you feel for your score by the judges on this pitch (versus feeling like ChatGPT is responsible)?” on a scale from 0 (*ChatGPT is 100% Responsible*) to 100 (*I am 100% Responsible*), and “How much **creativity** do you feel you used for this pitch?” on a scale from 0 (*No Creativity*) to 100 (*All of My Creativity*).

Controlling for Individual Differences. We also asked participants how they felt about the task and about how much experience they had using GenAI in order to control for these factors. Participants reported their comfort with writing (“How **comfortable** do you feel with writing tasks?”), a measure of task expertise, on a scale from 0% (*Not at All Comfortable*) to 100% (*Very Comfortable*). Participants then rated the difficulty of the task (“How **difficult** was this task for you?”) on a scale from 0 (*Not at All Difficult*) to 100 (*Very Difficult*), their experience with GenAI (“How much **experience** do you have using ChatGPT in general?”) using five radio buttons from *No Experience* to *Extensive Experience*. They also answered the manipulation check for expertise, “How **knowledgeable** do you feel about the topic assigned to you?” on a scale from 0 (*Not at all Knowledgeable*) to 100 (*Very Knowledgeable*).

Following data collection, each participant’s initial and final pitches were judged by a separate sample of judges recruited from Prolific. Judges were blind to the condition used to create the proposal and scored them from 0 to 100 on four criteria (see Table 1). For each pitch, we created the **average judge score** by averaging the scores across the three judges and rescaling them to a scale from 0 to 8. **Total time** was calculated using the time it took to complete the survey.

In both Experiment 1 and 2, we constructed a productivity measure that combines two critical dimensions of work performance: quality and time. While prior research examined quality and time as separate performance indicators (Chew 1988, Martin 2015, Noy and Zhang 2023), we combined them into a single metric in order to capture the fundamental trade-off people face when collaborating with GenAI. Doing so allowed us to make inferences about productivity as a single ratio.

We calculated the productivity score by dividing the average judge score on the final pitch and dividing it by the total time:

$$\text{Productivity} = \frac{\text{Average Judge Score}}{\text{Total Time}}$$

Finally, we calculated the number of words in each podcast pitch and the similarity between the initial and final versions using text similarity measures.

1.2. Analysis Strategy

We ran t-tests and regressions for all variables of interest; regressions are in the main text, t-tests are in the Appendix A, and all preregistered models are on OSF. Although our topic expertise manipulation was successful, we did not find any significant interactions between expertise and GenAI collaboration strategy. Thus, we report the OLS regressions without interactions.¹ When we found non-significance between our conditions, we report results of the Bayesian t-tests (with a .707 prior) to assess the likelihood of the null relative to the alternative (BF_{01}).

¹ Participants in the high-expertise condition rated themselves as more knowledgeable on the topic they were assigned than those in the low-expertise condition (Mean Difference = -33.6, $t(543.27) = -16.09$, $p < .001$; Cohen’s $d = -1.38$). Moreover, we did not include knowledgeable as a covariate in our models as it was highly correlated with our Expertise manipulation ($r = .55$). Finally, we included experience with GenAI as a covariate in all of our models, without an a priori hypothesis.

1.3. Results & Discussion

We report our findings as follows. First, we examined which GenAI collaboration strategy was optimal for productivity. Second, we analyzed which GenAI collaboration strategy was optimal for the psychological experience of work. Third, we examined the mechanisms driving productivity by analyzing total time and average judge score.

Productivity. We examined participants' productivity, as measured by the average judge score divided by total time (see Figure 1). Surprisingly, we found that the collaboration strategies were similarly productive (as were experts and non-experts). It did not matter whether GenAI was used as a drafter (see Table 2), $b = .08, t(595) = -1.34, p = .180; \beta = -.11, BF_{01} = 1.35$ or whether the topic expertise was high ($b = -.07, t(595) = 1.57, p = .116; \beta = 0.06, BF_{01} = 5.12$).

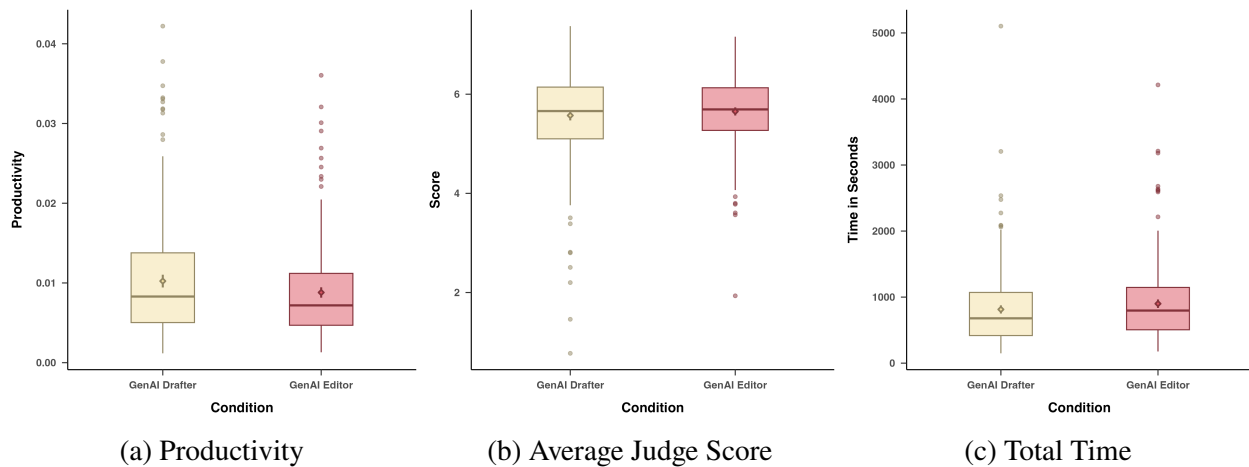


Figure 1 Exp. 1: Productivity, Average Judge Score, and Total Time.

To further understand productivity, we examined its components, average judge score, and total time separately. Table 2 shows that using GenAI as a drafter decreased total time ($b = -.11, t(595) = -2.19, p = .029; \beta = -0.09$) but that the average judge score was similar between the GenAI collaboration strategies ($b = -.10, t(595) = -1.43, p = 0.154; \beta = -0.06, BF_{01} = 5.25$). Expertise in the topic neither influenced average judge score ($b = .01, t(595) = 0.08, p = 0.936; \beta = 6.52e - 03, BF_{01} = 10.90$) nor the total time ($b = .08, t(595) = 1.55, p = 0.123; \beta = 0.12, BF_{01} = 3.87$). Although people spent less time creating podcast pitches when they used GenAI to draft, the time savings were not substantial enough to affect productivity, controlling for task difficulty, comfort with writing, and experience with GenAI. To unpack this result, we examined drivers of quality and total time in detail in the section “Unpacking Quality and Total Time.”

Psychological Experience of Work. Although the GenAI collaboration strategies were similarly productive, the strategies produced marked differences for the psychological experience of work (see Figure 2). Using GenAI as a drafter rather than as an editor resulted in lower feelings of control (see Table 3, $b = -7.49,$

Table 2 Exp. 1: OLS Regression Predicting Productivity

	<i>Dependent variable:</i>		
	Log(Productivity)	Average Judge Score	Log(Total Time)
	(1)	(2)	(3)
Intercept	-4.54*** (.14)	5.86*** (.18)	6.30*** (.13)
Expertise - High	-.07 (.05)	.01 (.07)	.08 (.05)
GenAI Role - Drafter	.08 (.05)	-.10 (.07)	-.11* (.05)
Comfort w/ Writing	-.003* (.001)	-.001 (.002)	.003* (.001)
Task Difficulty	-.004*** (.001)	-.003* (.001)	.003** (.001)
Experience w/ GenAI	.01 (.03)	-.01 (.03)	-.01 (.02)
Observations	601	601	601
R ²	.03	.01	.03
Adjusted R ²	.02	.001	.03
Residual Std. Error (df = 595)	.64	.81	.60
F Statistic (df = 5; 595)	4.01**	1.10	4.31***

Notes. Reference categories: Low Expertise, GenAI Role - Editor (participant drafts, GenAI edits).

Both productivity and total time were log-transformed to address right skew.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

$t(595) = -3.64$, $p < .001$; $\beta = -0.14$), responsibility ($b = -7.74$, $t(595) = -3.80$, $p < .001$; $\beta = -0.14$), and creativity ($b = -17.11$, $t(595) = -8.58$, $p < .001$; $\beta = -0.30$), controlling for comfort with writing, task difficulty, and experience with GenAI. Using GenAI as a drafter degraded people's psychological experience of work without productivity benefits. Relative to using GenAI to draft, using GenAI to edit allowed people to maintain their control, responsibility, and creativity.

Additionally, participants with higher topic expertise reported greater feelings of control overall (see Table 3, $b = 5.44$, $t(595) = -3.64$, $p < .01$; $\beta = -0.14$). Yet, the positive impact of topic expertise was not strong enough to overcome the negative impact of using GenAI as a drafter on the reported feelings of control.

Unpacking Quality and Total Time. Contrary to our pre-registered predictions, the different GenAI collaboration strategies were similarly productive. In order to understand why, we examined quality and total time.

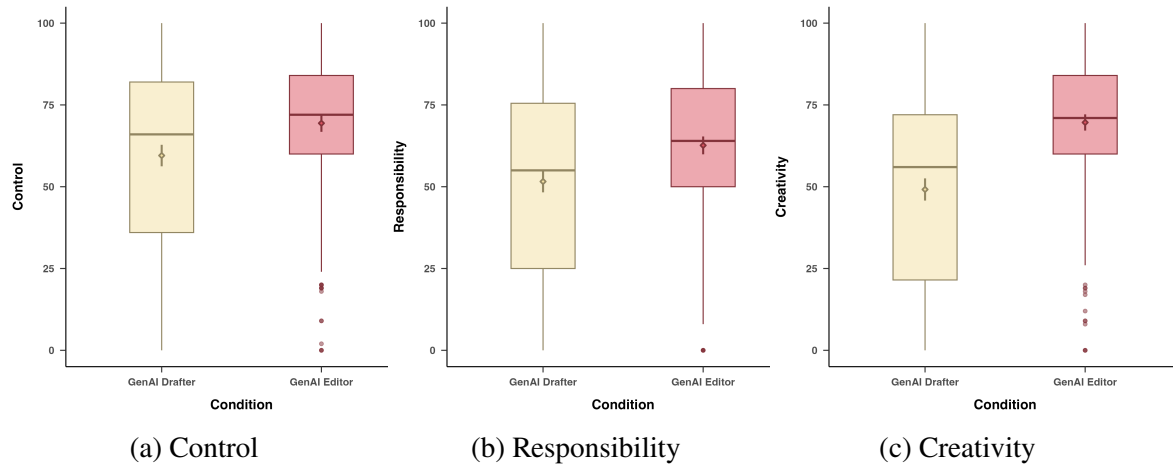


Figure 2 Exp. 1: Psychological Experience of Work

Table 3 Exp. 1: OLS Regression Predicting Psychological Experience of Work

	<i>Dependent variable:</i>		
	Control	Responsibility	Creativity
	(1)	(2)	(3)
Intercept	27.89*** (5.54)	7.24 (5.48)	15.98** (5.37)
Expertise - High	5.44** (2.03)	.31 (2.01)	.55 (1.97)
GenAI Role - Drafter	-7.49*** (2.06)	-7.74*** (2.04)	-17.11*** (1.99)
Comfort w/ Writing	.39*** (.05)	.45*** (.05)	.50*** (.05)
Task Difficulty	.09* (.04)	.26*** (.04)	.23*** (.04)
Experience w/ GenAI	1.97 (1.01)	3.32*** (1.00)	2.02* (.98)
Observations	601	601	601
R ²	.15	.20	.28
Adjusted R ²	.14	.19	.28
Residual Std. Error (df = 595)	24.90	24.64	24.13
F Statistic (df = 5; 595)	20.67***	29.49***	46.52***

Notes. Reference categories: Low Expertise, GenAI Editor (participant drafts, GenAI edits), *p<0.05;

p<0.01; *p<0.001.

Quality. One possible explanation for why the different GenAI collaboration strategies produced similar quality deliverables is that judges were insensitive to quality differences. However, this explanation is unlikely given that the average judge score was responsive to task difficulty (Table 2, $b = -.003$, $t(595) = -1.99$, $p < 0.05$; $\beta = -0.09$), suggesting that judges differentiated across pitches.

A more plausible explanation is that both strategies involved using GenAI, leading to final versions that were similar in quality. Prior work comparing the presence versus absence of GenAI found large differences in quality (Noy and Zhang 2023). In our experiment, GenAI's involvement in both collaboration strategies dominated the final output, additionally masking differences across topic expertise.

In order to understand the contribution of GenAI relative to people, we examined the number of words (Table 4, column 1) and editing patterns measured by changes in the number of words and text similarity between initial and final pitches (Table 4, columns 2 and 3). Using GenAI as a drafter produced shorter final versions compared to using GenAI as an editor, controlling for comfort with writing, task difficulty and experience with GenAI (Table 4(1), $b = -17.42$, $t(595) = -4.43$, $p < .001$; $\beta = -0.18$). Furthermore, GenAI added more words when editing than people did (Table 4(2), $b = -125.75$, $t(595) = -21.91$, $p < .001$; $\beta = -0.67$) and the initial and final versions were more dissimilar² when GenAI edited (Table 4(3), $b = -.22$, $t(595) = -17.53$, $p < .001$; $\beta = -0.59$).³

When we calculated the difference in length between the initial and final deliverables, GenAI added an average of 99 words to the drafts produced by people. In contrast, people deleted an average of 23 words from the draft produced by GenAI. While a longer final deliverable resulted in a higher average judge score for the sample as a whole (Table 4(5), $b = .01$, $t(592) = 8.27$, $p < .001$; $\beta = .33$), the amount of edits people made did not meaningfully affect the average judge score in comparison to the edits made by GenAI (Table 4(5), $b = .01$, $t(592) = -0.13$, $p = .894$; $\beta = -0.01$). However, when we measured the improvement in the average judge score from the initial to the final deliverable, people's edits did not improve the proposal's average judge score as much as GenAI's edits did (Table 4(4), $b = -.94$, $t(596) = 9.84$, $p < .001$; $\beta = 0.75$).

Total Time. Regardless of the workflow within the GenAI collaboration strategy, i.e., whether GenAI entered the workflow as a drafter or editor, the contribution of GenAI seemed to dominate the final output. Using GenAI as a drafter reduced total time, which could have been driven by a minimal amount of time spent reviewing the GenAI draft or by limited editing of content. We addressed this in Experiment 2 by examining the time spent on various components of the task: writing, reviewing, and editing.

Robustness Checks. We checked the robustness of our results by examining where participants may have misinterpreted the task (for example, entering prompts rather than submitting a podcast pitch draft in the GenAI edit condition). To do this, we removed final deliverables that were unusually short or unusually

² A cosine difference equal to 1 means completely different texts, 0 means the same text.

³ We calculated additional measures of text similarity (Jaccard and Sorensen-Dice) and they showed the same pattern.

Table 4 Exp. 1: OLS Regression of Text Length, Similarity, and Average Judge Score Differences

	<i>Dependent variable:</i>				
	# Words (1)	Δ Words (2)	Cosine Diff. (3)	Δ Score (4)	Avg. Judge Score (5)
Intercept	258.20*** (10.60)	155.29*** (15.46)	.35*** (.03)	1.45*** (.26)	4.38*** (.25)
Expertise - High	3.03 (3.89)	3.12 (5.67)	.001 (.01)	.04 (.09)	-.01 (.06)
GenAI Role - Drafter	-17.42*** (3.94)	-125.75*** (5.74)	-.22*** (.01)	-.94*** (.10)	.01 (.06)
Comfort w/ Writing	-.03 (.10)	-.27 (.14)	.0004 (.0003)	-.005 (.002)	-.001 (.002)
Task Difficulty	-.18* (.08)	-.45*** (.11)	.0000 (.0002)	-.01** (.002)	-.001 (.001)
Experience w/ GenAI	-.61 (1.93)	-5.68* (2.82)	-.02* (.01)	-.05 (.05)	-.01 (.03)
Total Duration					-.0001 (.0001)
# Words					.01*** (.001)
Self Rating					.03 (.02)
Observations	601	601	601	601	601
R ²	.04	.45	.35	.15	.13
Adjusted R ²	.03	.45	.35	.14	.12
Residual Std. Error	47.62	69.45	.15	1.15	.76
F Statistic	4.90***	97.64***	64.90***	20.41***	10.89***

Notes. Reference categories: Low Expertise, GenAI Editor (participant drafts, GenAI edits), *p<0.05;

p<0.01; *p<0.001.

long, i.e., those in the top/bottom 2.5% of the values (excluding a total of 5% of the data). The excluded deliverables were also checked to ascertain that the thresholds we used correctly identified misinterpretation by the participants. This robustness check is presented in Appendix A. Our results held with this exclusion.

Discussion. We identified the optimal GenAI collaboration strategy between using GenAI to draft versus edit (RQ1): using GenAI to edit. Using GenAI as a drafter degraded the psychological experience of work without productivity benefits. Counter to our predictions, neither GenAI collaboration strategy nor topic

expertise influenced productivity. While using GenAI as a drafter decreased total time, it did not improve quality and was not enough to improve productivity.

The asymmetrical effect of strategy on productivity and psychology raises a critical question for organizations integrating GenAI into workflows: which GenAI collaboration strategy will people actually choose? If they gravitate toward using GenAI to draft despite its psychological costs, perhaps due to perceived ease or novelty, organizations may face long-term challenges in employee engagement and sustainable GenAI adoption. In Experiment 2, we directly examine preferences when people choose between these collaboration strategies.

Uncovering the mechanics of the human-GenAI collaboration (RQ4) helped explain why the collaboration strategies produced similar levels of productivity. First, regardless of when GenAI entered the workflow (as a drafter or editor), it added a substantial amount of text to the deliverable. Second, while people deleted text from drafts generated by GenAI, the magnitude of these changes did not meaningfully affect the average judge score. Consequently, although judges were responsive to the length of the deliverable, the final outputs received similar average judge scores. Additionally, as one might expect, the improvement in quality from the initial to the final version of the deliverable was greater when GenAI edited due to the magnitude of text GenAI added. Moreover, higher topic expertise was associated with greater feelings of control, regardless of the strategy. Even if there were differences between experts and non-experts in their strategy use, the presence of GenAI itself may have dominated quality.

One alternative explanation for our results in Experiment 1 is that productivity benefits from collaboration strategies were not different because participants were randomly assigned to a GenAI collaboration strategy and thus were not necessarily working with the strategy that was natural or beneficial to the way they normally work. We addressed this possibility in Experiment 2 by asking participants to select their GenAI collaboration strategy. It is also possible that the participants in Experiment 1 tried to optimize for time even though they were monetarily incentivized to produce a high quality deliverable. We addressed this in Experiment 2 by recruiting a sample who could potentially use the deliverable they created in the experiment for their job. Finally, Experiment 2 examines the mechanics of the GenAI-human collaboration by unpacking timing for the different collaboration stages.

2. Experiment 2: Lesson Plan

The primary goals of Experiment 2 were to test for the optimal GenAI collaboration strategy (RQ1), to test if people have preferences for their GenAI collaboration strategy (RQ2), to identify whether people select the optimal strategy (RQ3), and to test the mechanisms of GenAI-human collaboration (RQ4). In contrast to Experiment 1, where participants were assigned to a GenAI collaboration strategy, participants in Experiment 2 chose their strategy. Thus, examining strategy choices (rather than comparing the presence versus absence of GenAI as in earlier work), allowed us to measure: 1) preferences for collaboration with GenAI, 2) whether

those preferences were optimal for maximizing productivity and the psychological experience of work, and 3) the individual differences that drove these preferences. Our primary outcome measures were: choice, productivity, and the psychological experience of work (perceived control, responsibility, and creativity).

We chose an ecologically valid task in order to motivate participants to optimize their workflow for both quality and time. We recruited teachers, who taught grades 4 to 12, in order to create a lesson plan that they could use in their classes. We expected that the potential to use the task for work (in addition to the monetary incentive) would motivate participants to create a high-quality plan. In order to leverage this increased ecological validity, we extended our psychological metrics to perceptions of the strategies and expectations about using them for their job: frustration with the process, motivation to use the strategy they chose in the future, the likelihood of using their lesson plan for work, and whether they had high enough reputational concerns with the strategy to change their choice.

We compared the mechanics of the collaboration between the collaboration strategies. First, we tested the mechanism driving the relationship between the GenAI collaboration strategy and quality by parsing time spent on different task components within the workflow (writing, reviewing, editing). Second, we examined the number of words as a proxy for effort. Third, we tested whether GenAI increased the number and diversity of ideas by measuring the number of lesson objectives.

2.1. Method

The final sample included 698 participants who submitted a lesson plan (51.6% women, 46.7% men, and 1.7% other; median age = 36). We recruited teachers on Prolific for a study about creating a lesson plan. Participants were compensated \$1.50 for a ten-minute survey. As in Experiment 1, participants could receive a bonus of up to an additional \$1.50, based on the average judge score by three judges. Judges from Prolific ($N = 1396$) each scored three lesson plans.

Design. We employed a two-cell between-subjects experimental design (No GenAI vs. Choose GenAI Collaboration Strategy). Participants were randomly assigned to either the control condition ($n = 231$; 33.1%) or the GenAI collaboration condition ($n = 467$; 66.9%). Those in the control condition completed the lesson plans without GenAI assistance, providing us with a benchmark comparison. Those in the GenAI collaboration condition selected a strategy, using GenAI as a drafter or editor.

Procedure and Materials. To qualify for the experiment, participants reported the grade level they taught (4th through 12th grade). They read about the task, creating a short lesson plan for a 45-50 minute class. They read, “We want this to be as helpful to you in the classroom as possible, so please choose a topic/subject you would be likely to teach.” Then, as in Experiment 1, they read about the monetary incentive based on their average judge score. They read the four criteria that judges would use (see Table 5) and the bonus details. On the next page, they reported their grade level, their subject from a list (e.g., history, math, science, or other), and entered their topic for the class session into an open-ended text response.

Table 5 Evaluation criteria in Experiment 2

Criteria	Explanation
Objectives	Does the lesson have clear objectives and activities that match those objectives?
Content Quality	Does the lesson include key concepts and is appropriate for the grade level?
Feasibility	How realistic would the lesson be to teach in an actual classroom?
Student Appeal	How engaging and inclusive the lesson would be for students.

Participants in the control condition were instructed to write their lesson plan. They did not read any information about GenAI assistance. Participants in the GenAI collaboration condition chose between two GenAI collaboration strategies: using GenAI to edit and using GenAI to draft.

In order to maintain a clean experimental design, we allowed participants a one-shot interaction with GenAI and did not allow them to iterate beyond that. Participants who chose to use GenAI as an editor were asked to draft their own lesson plan. In both of the GenAI collaboration strategy conditions, participants read what was sent to GenAI: their draft, grade level, subject, and topic, and the prompt to GenAI which included the judging criteria. Participants were shown GenAI's edited plan and, without further modification, submitted that as their final lesson plan. Participants who chose to have GenAI draft read GenAI's draft and made as many or as few edits as they wanted in order to produce a final plan, creating comparable conditions. In both conditions, participants estimated the average judge score for their initial and final plans and specified how many lesson objectives their plan had. Next, all participants answered questions about their psychological experience of work, individual differences, intentions to utilize both the strategy and the lesson plan beyond the experiment, and demographic information, as detailed below.

Psychological Experience of Work. Participants rated their psychological experience of work on the same three key dimensions as Experiment 1 (Control, Responsibility, and Creativity) on a scale from 0 (*not at all*) to 100 (*extremely*). In order to measure how GenAI influenced how people engaged with the work, we asked participants about their intentions to adopt the strategy and to utilize the lesson plan because they could realistically use the work they created during the experiment for their job. This task allowed us to then capture participants' motivation to utilize the strategy and the plan itself beyond the scope of the experiment. Specifically, we asked "How **motivated** are you to generate lesson plans using the strategy you employed during your session?" and, "How likely are you to **use** the final lesson plan you generated in this session in your class?" Participants answered both questions using a 0% (*motivated/likely*) to 100% (*motivated/likely*) scale. We also asked "How **frustrated** were you with this process?" on a 0 (*not at all*) to 100 (*extremely*) scale. Furthermore, in order to understand whether people's preferences for GenAI collaboration change when considering their reputation, we asked them, "If the **judge knew** your strategy, would you have changed your mind?" and they chose Yes, No, or Unsure.

Individual Differences Driving Psychological Outcomes and Choice. We aimed to better understand the mechanisms driving people's choices and tested not only whether individual differences interacted with the GenAI collaboration strategy and the psychological experience of work but also whether individual

differences were associated with strategy choice itself. As in Experiment 1, participants reported their task difficulty, comfort with the task, and experience with GenAI. Additionally, in order to measure expertise, we asked the participants to report their years of classroom teaching experience (tenure).

Mechanics of Collaboration. In order to understand how people spent time when collaborating with GenAI, we parsed time on task into: time spent writing, reviewing, and editing. We then used this total time-on-task measure for each condition to calculate productivity in the same way as we did in Experiment 1. Additionally, we calculated the number of words for each plan and the text similarity between the initial and final versions.

Following data collection, each initial and final lesson plan was independently evaluated by three judges recruited from Prolific who were blind to the condition used to create each lesson plan. For each of the four criteria in Table 5, judges scored lesson plans on a scale from 0 (*not at all*) to 100 (*extremely*). These were rescaled to a 0 to 8 scale in order to match the bonus payment scale.

2.2. Analysis Strategy

Following our pre-registration, for each outcome, we ran group comparisons between all conditions (i.e., t-tests, z-tests). For each t-test where we did not find significance, we ran a Bayesian t-test (using a .707 prior) to investigate the probability of the null versus the alternative hypothesis. For each of our outcomes, we ran a General Linear Model (i.e., OLS, Logistic Regression when appropriate) including the covariates: Task Comfort, Task Difficulty, Experience with GenAI, Tenure, and STEM. Our main comparisons were outcomes when people used GenAI compared to when they did not (the control condition) and comparing the GenAI collaboration strategies to each other (when they used GenAI to draft versus edit). We report additional models on OSF (e.g., models including only the experimental conditions and Tenure, models including interactions and additional covariates such as subject designation and grade level). Tenure neither affected our outcomes as a main effect nor as an interaction with the experimental conditions.

2.3. Results & Discussion

We report our findings as follows. First, we examined which GenAI collaboration strategy people preferred (RQ 2). Second, we examined productivity. Third, we examined how the GenAI collaboration strategies differentially influenced the psychological experience of work (RQ1 and RQ 3). Fourth, we examined the mechanics of the collaboration by testing how each strategy differed in quality and time on task (RQ4).

Strategy choice. While 61.7% of participants chose to have GenAI draft their plan, only 38.3% chose to have GenAI edit ($X^2 = 49.95$, $df = 1$, $p < .001$), suggesting a strong preference to use GenAI to draft. We found no significant individual differences that predicted the GenAI collaboration strategy choice (e.g., tenure, gender, comfort with the task).

Productivity. We examined people's productivity, as measured by the average judge score divided by participants' time on task (see Figure 3 for boxplots). Participant time on task was operationalized as:

- *Control Condition*: writing time
- *GenAI Drafter Condition*: reviewing and editing time
- *GenAI Editor Condition*: writing and reviewing time

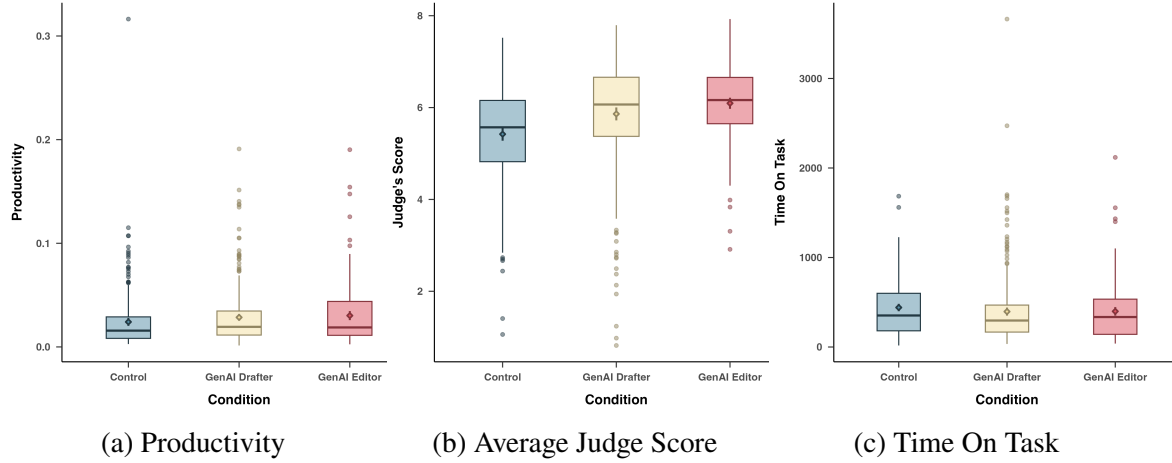


Figure 3 Exp. 2: Productivity, Average Judge Score, and Time on Task

As predicted, people who used GenAI were more productive than those who did not use GenAI (Table 6(1), $b = .22$, $t(681) = 2.49$, $p = 0.013$, $\beta = .12$), consistent with prior work (Noy and Zhang 2023, Brynjolfsson et al. 2025). Counter to our prediction, even with a highly motivated sample, teachers were similarly productive regardless of whether GenAI drafted or edited lesson plans ($b = -.04$, $t(681) = -.42$, $p = .673$; $\beta = -.02$, $BF_{01} = 7.71$). We found that the more experience teachers had using GenAI, the higher their productivity ($b = .07$, $t(681) = 2.24$, $p = .025$; $\beta = 0.09$). No other measures explained productivity.

To better understand productivity, we examined its components: quality and time on task. Table 6 suggests that *quality*, and not time, drove differences in productivity between the presence and absence of GenAI. The productivity gain we found due to the presence of GenAI was driven by the increase in average judge score (Table 6(2), $b = .62$, $t(680) = 5.70$, $p < .001$; $\beta = .26$). We identified the mechanism driving the variation in average judge scores: the number of words. When we included the number of words, all other significant predictors (using GenAI and tenure) became non-significant. Additionally, the number of words fully mediated the effect of GenAI collaboration strategy. This result suggests that the verbosity of GenAI drove quality (see Appendix B for mediation analysis).

Counter to our prediction, teachers did not save time by using GenAI relative to not using it (Table 6(3), $b = -.09$, $t(686) = -1.05$, $p = .295$; $\beta = -.05$). When considering collaboration strategies, one might expect comparable productivity between these strategies. Here, it was plausible that using GenAI to edit might increase quality due to teachers' high-quality lesson plan drafts, while taking longer. However, unexpectedly, the two collaboration strategies produced similar average judge scores ($b = -.20$, $t(681) = -1.93$, $p = 0.053$; $\beta = -.09$, $BF_{01} = .81$) and similar amount of time on task ($b = -.03$, $t(686) = -.35$, $p = .725$; $\beta = -0.02$,

Table 6 Exp. 2: OLS Regression Predicting Productivity

	<i>Dependent variable:</i>		
	log(Productivity) (1)	Average Judge Score (2)	log(Time On Task) (3)
Intercept	−4.06*** (.23)	5.24*** (.29)	5.69*** (.22)
GenAI Used	.22* (.09)	.62*** (.11)	−.09 (.08)
GenAI Drafter	−.04 (.08)	−.20 (.10)	−.03 (.08)
Tenure	−.01 (.004)	−.02** (.01)	.004 (.004)
Task Comfort	−.002 (.002)	.004 (.003)	.003 (.002)
Task Difficulty	−.002 (.001)	−.003 (.002)	.001 (.001)
Experience w/ GenAI	.07* (.03)	.03 (.04)	−.07* (.03)
STEM	−.07 (.07)	−.02 (.08)	.07 (.06)
Observations	689	689	694
R ²	.03	.07	.02
Adjusted R ²	.02	.07	.01
Residual Std. Error	.86 (df = 681)	1.08 (df = 681)	.82 (df = 686)
F Statistic	3.01** (df = 7; 681)	7.88*** (df = 7; 681)	1.70 (df = 7; 686)

Note:

*p<0.05; **p<0.01; ***p<0.001

The regression models presented are ordinal linear models. The “AI used” coefficient represents the effect of using GenAI, regardless of GenAI collaboration strategy. The “GenAI drafter” coefficient represents the additional effect of using GenAI to draft (beyond the baseline effect of using GenAI to edit). In the “Productivity” column, we compared productivity based on the average judge score for the final plans between conditions.

$BF_{01} = 2.23$). In order to understand this result, we further examined drivers of time on task and quality in the section “Unpacking Time and Quality,” after reporting the psychological experience of work results. Taken together, our results suggest that using GenAI influences productivity, but there is no additional effect on productivity based on the strategy, even when people choose it.

Psychological Experience of Work. Using GenAI was detrimental to people's experience of work (see Figure 4 for boxplots). They experienced less control and responsibility when using GenAI (see Table 7, $b = -10.73$, $t(686) = -6.40$, $p < .001$; $\beta = -0.28$ and $b = -10.13$, $t(686) = -5.37$, $p < .001$; $\beta = -0.23$, respectively). GenAI Collaboration did not impact people's feelings of creativity ($b = -3.21$, $t(686) = -1.56$, $p = .118$; $\beta = -0.07$).

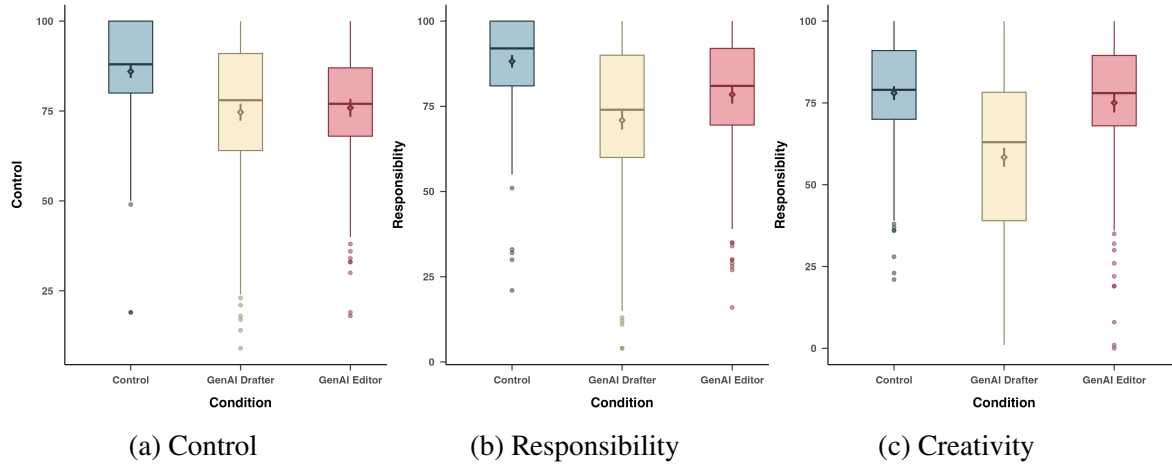


Figure 4 Exp. 2: Psychological Experience of Work

Contingent on using the tool, leveraging GenAI as a drafter further damaged people's experience of work. They felt less responsible and creative when using GenAI to draft than edit ($b = -7.22$, $t(686) = -4.04$, $p < 0.001$, $\beta = -.17$ and $b = -16.38$, $t(686) = -8.42$, $p < 0.001$, $\beta = -.35$, respectively). People experienced similar feelings of control between the two GenAI collaboration strategies ($b = -.73$, $t(686) = -.46$, $p = .646$, $\beta = -.02$; from t-test $BF_{01} = 7.61$). These results suggest that there are psychological costs to leveraging GenAI and that using GenAI to draft even further degrades feelings of responsibility and creativity.

Individual Differences Driving Psychological Outcomes. Two drivers increased positive psychological outcomes: how comfortable people were creating the lesson plan and their previous experience with GenAI. Specifically, the more comfortable people were completing the task, the higher their perceived control ($b = .37$, $t(686) = 7.95$, $p < 0.001$, $\beta = .31$), responsibility ($b = .37$, $t(684) = 7.07$, $p < 0.001$, $\beta = .27$), and creativity ($b = .42$, $t(684) = 7.31$, $p < 0.001$, $\beta = .28$). Similarly, the more experience people had using GenAI, the more control ($b = 1.56$, $t(684) = -0.24$, $p = 0.012$, $\beta = .09$), responsibility ($b = 1.37$, $t(686) = 1.96$, $p = 0.049$, $\beta = .07$), and creativity they felt ($b = .42$, $t(684) = 2.56$, $p = 0.006$, $\beta = .28$). Comfort with the task and experience with GenAI appeared to buffer the cost of GenAI on people's psychology. Difficulty with the task worked somewhat differently; the more difficult the task felt, the more creative people felt ($b = .16$, $t(684) = 4.73$, $p < 0.001$, $\beta = .17$).

Task Engagement. In order to examine the implications of these strategies for our participants beyond the experiment, we asked them about their intentions to use the GenAI collaboration strategies in the future.

Table 7 Exp. 2: OLS Regression Predicting Psychological Variables

	<i>Dependent variable:</i>		
	Control (1)	Responsibility (2)	Creativity (3)
Intercept	50.23*** (4.47)	50.90*** (5.04)	30.86*** (5.48)
GenAI Used	-10.73*** (1.68)	-10.13*** (1.89)	-3.21 (2.05)
GenAI Drafter	-.73 (1.59)	-7.22*** (1.79)	-16.38*** (1.95)
Tenure	.06 (.09)	.08 (.10)	.12 (.10)
Task Comfort	.37*** (.05)	.37*** (.05)	.42*** (.06)
Task Difficulty	-.01 (.03)	.04 (.03)	.16*** (.03)
Experience w/ GenAI	1.56* (.62)	1.37* (.70)	2.05** (.76)
STEM	.44 (1.30)	2.36 (1.46)	1.25 (1.59)
Observations	694	694	694
R ²	.20	.21	.24
Adjusted R ²	.19	.20	.23
Residual Std. Error (df = 686)	16.52	18.60	20.25
F Statistic (df = 7; 686)	24.06***	26.08***	31.34***

Note: *p<0.05; **p<0.01; ***p<0.001

People were similarly motivated to continue using their assigned or chosen strategy in creating the plan, regardless of whether they used GenAI or how they used it (see Table 8, $b = -.57$, $t(686) = -0.32$, $p = .746$, $\beta = -.01$ and $b = -2.09$, $p = .213$, $\beta = -.05$). Both those who felt more comfortable with the task and those who had more experience using GenAI reported stronger motivation to continue using their strategy ($b = .45$, $t(686) = 9.08$, $p < 0.001$, $\beta = .36$ and $b = 1.93$, $t(686) = 2.95$, $p < 0.01$, $\beta = .11$).

Importantly, teachers reported that they were less likely to use their lesson plan in the classroom when they used GenAI, compared to when they did not ($b = -3.92$, $t(686) = -2.12$, $p = .03$, $\beta = -.09$). This effect held when controlling for teaching experience and task comfort, both of which positively affected the likelihood of their future use ($b = .25$, $t(686) = 2.60$, $p < .01$; $\beta = 0.10$ and $b = .36$, $t(686) = 6.98$, $p < .001$; $\beta = 0.28$).

Table 8 Exp. 2: OLS Regression Predicting Motivation to Use GenAI Collaboration Strategy and Likelihood of Using Lesson Plan

	<i>Dependent variable:</i>	
	Motivation to Use Strategy	Likelihood of Using Lesson Plan
Intercept	34.62*** (4.73)	48.56*** (4.94)
GenAI Used	-.57 (1.77)	-3.92* (1.85)
GenAI Drafter	-2.09 (1.68)	-3.25 (1.75)
Tenure	.17 (.09)	.25** (.09)
Task Comfort	.45*** (.05)	.36*** (.05)
Task Difficulty	.02 (.03)	-.01 (.03)
Experience w/ GenAI	1.93** (.65)	.87 (.68)
STEM	2.00 (1.37)	2.14 (1.43)
Observations	694	694
R ²	.17	.14
Adjusted R ²	.16	.13
Residual Std. Error	17.49 (df = 686)	18.24 (df = 686)
F Statistic	19.69*** (df = 7; 686)	16.21*** (df = 7; 686)
<i>Note:</i> *p<0.05; **p<0.01; ***p<0.001		

How frustrated did teachers feel? Using GenAI to edit resulted in feeling less frustrated compared to the control (difference = 5.24, $t(363.42) = 2.15$, $p = 0.032$; Cohen's $d = .23$, $BF_{01} = 1.03$, see Appendix B). People felt similarly frustrated with using GenAI to edit versus to draft (difference = 1.01, $t(344.35) = .44$, $p = .659$; Cohen's $d = 0.05$, $BF_{01} = 8.28$). Editing helped preserve the benefits of positive psychological experience of work compared to not using GenAI, without a cost to frustration, compared with drafting. Furthermore, the reputational concerns (whether the judges knew of the GenAI collaboration strategy used) were not strong enough to change the strategy choice (Proportion Z-test; difference = 1.58, $\%_{GenAIDrafter} = 15.6\%$, $\%_{GenAIEDitor} = 14.0$ $z(1) = .110$, $p = .74$).

Unpacking Time and Quality. To understand why GenAI collaboration strategies produced equivalent productivity, we first examined the time teachers spent on the task components and then unpacked quality.

Time on Task Components. We deconstructed time into three components: writing, reviewing GenAI output, and editing GenAI output (see Table 9). Two observations explained why the total time remained constant between the collaboration strategies (see Table 6). First, people who chose to have GenAI draft the plan spent substantial time reviewing the GenAI draft before editing it, which was equivalent to the time people spent reviewing when they chose to have GenAI edit. Second, those who chose to have GenAI edit, knowing that GenAI would edit their work, spent 5.62 minutes writing (less than those without GenAI). However, they spent time reviewing GenAI's edits for an additional 1.02 minutes, ultimately, losing out on any time saved.

This review time partially offset the time saved in both collaboration strategies as teachers did not write from scratch and they provided a shorter draft when expecting GenAI assistance. The review step, where people read through the GenAI output in both collaboration strategies, was necessary but not instantaneous. Together, the time needed to review and edit the draft when using GenAI to draft and the writing and reviewing effort when using GenAI to edit produced comparable total time on task across all conditions. These results provide greater insight into how people allocate effort when collaborating with GenAI. Counter to our prediction, using GenAI to draft did not necessarily save people time relative to using GenAI to edit and thus did not translate into productivity benefits.

Condition	Writing	Reviewing GenAI Output	Editing GenAI Output	Total
	M(SD)	M(SD)	M(SD)	M(SD)
Control	7.35(5.31)			7.35(5.31)
GenAI Drafter		1.43(2.04)	5.16(5.50)	6.59(6.45)
GenAI Editor	5.62(5.09)	1.02(1.01)		6.64(5.36)

Note: Participants who had GenAI Edit reviewed the GenAI output in order to forecast the average judge score.

Quality. Table 6 shows that GenAI collaboration, relative to not using GenAI, improved productivity by improving quality. Moreover, the collaboration strategies produced similar quality lesson plans. In order to test the drivers of improvement between the control condition and the collaboration strategies, we examined three dimensions: 1) text length, 2) the number of learning objectives, and 3) editing patterns measured by text similarity.

Why GenAI Collaboration Improves Quality. Compared to people who worked without GenAI assistance, those using GenAI (regardless of strategy) produced lesson plans with 206 more words (see Table 10(1), $b = 206.55$, $t(686) = 13.10$, $p < .001$; $\beta = 0.54$). Longer lesson plans translated directly into higher average judge scores (Table 10(6), $b = .003$, $t(666) = 11.42$, $p < .001$; $\beta = 0.46$). Additionally, when using GenAI

Table 10 Exp. 2: OLS Regression of Text Characteristics, Similarity, and Average Judge Score Differences

	<i>Dependent variable:</i>					
	# Words (1)	Objectives (2)	Δ Word (3)	Cosine Diff (4)	Δ Avg. Judge Score (5)	Avg. Judge Score (6)
Intercept	205.40*** (42.10)	2.05*** (.31)	184.88** (59.22)	.42*** (.07)	1.22* (.48)	4.74*** (.30)
GenAI Used	206.55*** (15.77)	.20 (.12)				.06 (.11)
GenAI Drafter	−26.11 (14.95)	.59*** (.11)	−262.74*** (17.72)	−.18*** (.02)	−1.02*** (.14)	−.14 (.10)
Tenure	−2.10** (.80)	−.02*** (.01)	−3.09** (1.19)	.002 (.001)	−.02* (.01)	−.01 (.01)
Task Comfort	.51 (.44)	.01* (.003)	.49 (.63)	−.001 (.001)	−.001 (.01)	.004 (.003)
Task Difficulty	−.33 (.26)	.01** (.002)	−.76* (.37)	.001 (.0004)	−.01** (.003)	−.002 (.002)
Experience w/ GenAI	8.09 (5.82)	.11** (.04)	8.04 (8.61)	.01 (.01)	.02 (.07)	.01 (.04)
STEM	−17.82 (12.20)	−.06 (.09)	−6.51 (17.62)	.04 (.02)	−.18 (.14)	.03 (.08)
Time on Task						−.0001 (.0001)
Words						.003*** (.0002)
Objectives						.001 (.03)
Self Score						−.03 (.03)
Observations	694	683	463	463	459	678
R ²	.27	.13	.35	.14	.13	.23
Adjusted R ²	.27	.12	.34	.13	.12	.22
Residual Std. Error	155.52	1.13	183.79	.22	1.49	.99
F Statistic	37.05***	14.33***	41.06***	12.70***	11.27***	17.86***

Note: *p<0.05; **p<0.01; ***p<0.001

to draft (versus to edit), the final lesson plan included 0.6 more learning objectives (Table 10(2), $b = .59$, $t(675) = 5.35$, $p < .001$; $\beta = 0.24$).

Why Collaboration Strategies Produce Similar Quality. Despite different workflows, the two strategies produced final lesson plans that were similar in length, controlling for all covariates (Table 10(1), $b = -26.11$,

$t(686) = -1.75, p = .081; \beta = -.07$). Thus, the strategies also produced a similar average judge score. While using GenAI to draft resulted in 0.6 more objectives, that did not make a practical difference.

In order to understand why the different strategies produced final plans that were similar in length, we compared the initial to the final lesson plans on length and similarity of text. Between the two GenAI collaboration strategies, editing behavior was asymmetric. Table 10(3) shows that when GenAI edited the person's draft, it added 185 more words, whereas when people edited the GenAI draft, they reduced the number of words by 80. These results suggest that when GenAI and people collaborate, GenAI adds a *greater* amount of text than people remove. Furthermore, in Table 10(4), the cosine difference⁴ between the final and initial lesson plans shows that people edited GenAI drafts less than GenAI edited people's drafts ($b = -.18, t(456) = -8.42, p < .001; \beta = -0.37$)⁵. This led to less improvement in the average judge score from the initial plan to the final plan when GenAI was used to draft versus edit ($b = -1.02, t(452) = -7.12, p < .001; \beta = -0.31$). Again, regardless of where GenAI entered the workflow, it acted as a hammer, overwhelmingly influencing outcomes.

Interestingly, people were accurate in predicting the value added from GenAI when they used GenAI to edit (prediction of Δ Average Judge Score=.73 versus actual Δ Average Judge Score=.78), however, they did not forecast that their edits would lead to a decrease in average judge score (prediction of Δ Average Judge Score=0.45 versus actual Δ Average Judge Score=-.32). Thus, people overestimated the value that they added within the workflow when GenAI drafted (difference=.82, $t(396.58) = 4.96, p < .001$; Cohen's $d = 0.50, BF_{01} = 1.22e^{-4}$, see Appendix B).

Robustness Checks. We checked the robustness of our results by examining where participants may have misinterpreted the task (for example, entering prompts rather than submitting a teaching plan in the GenAI edit condition). To do this, we removed final deliverables that were unusually short or unusually long, i.e., those in the top/bottom 2.5% of the values (excluding a total of 5% of the data). The excluded deliverables were also checked to confirm that the thresholds we used correctly identified misinterpretation by the participants. This robustness check is presented in Appendix B. Our results held with this exclusion.

Discussion. We identified the optimal GenAI collaboration strategy between drafting and editing: using GenAI to edit (RQ1), consistent with Experiment 1. Importantly, the majority of people chose the suboptimal GenAI collaboration strategy (RQ2, RQ3), which degraded their feelings of responsibility and creativity, without any added productivity benefits. Additionally, we identified factors that 1) improved the benefits of using GenAI and 2) buffered against its costs. Regardless of the GenAI collaboration strategy that people chose, those with GenAI experience were more productive. Moreover, people who were more comfortable with the task and people who had more experience with GenAI were buffered from the cost of GenAI. They

⁴ A cosine difference equal to 1 means completely different texts, 0 means the same text.

⁵ We calculated additional measures of text similarity (Jaccard and Sorensen-Dice), and they showed the same pattern.

felt more in control, responsible, and creative relative to those who felt less comfortable with the task and who had less experience.

Beyond the immediate effects of using GenAI, we were interested in understanding how people thought about using GenAI in their work after the experiment. Participants were comfortable continuing to utilize the same strategy they selected. Yet, they were less likely to actually use the lesson plan created in collaboration with GenAI compared with the plan created without GenAI.

These findings reveal an important trade-off for organizations. While GenAI collaboration improved immediate task productivity, people showed greater commitment to implementing outputs they created independently. This pattern suggests that foregoing GenAI assistance may offer long-term benefits, allowing individuals to build expertise and enhance ownership of the entire downstream process. In this case, teachers who developed their own lesson plans may have felt better prepared to deliver them effectively. Organizations must therefore weigh task-specific trade-offs when deciding whether to deploy GenAI support, balancing short-term productivity benefits against potential costs to worker engagement and long-term skill development.

Examining the mechanics of the collaboration (RQ4), we found that using GenAI increased the number of words, thus improving quality and productivity. Interestingly, the specific collaboration strategies were similar in their quality, time, and productivity. Counter to past literature, using GenAI did not save our participants time, suggesting that people put in similar effort whether they thought GenAI would edit or not and were willing to spend time reviewing and editing GenAI's draft.

Overall, the results from Experiment 2 revealed that while using GenAI as an editor was the optimal strategy for balancing both productivity and the psychology of work, the majority of people chose GenAI to draft. When editing, people refined text rather than generating it. In contrast, GenAI expanded and transformed text. GenAI added more words (185) than teachers removed (-80 words). GenAI's contribution, regardless of the sequencing within the workflow, dominated the final lesson plan quality. Including GenAI in the workflow created distance between workers and deliverables despite the improved quality and productivity.

3. General Discussion

Our research contributes to the new and growing literature on how people collaborate with GenAI by demonstrating that how people use GenAI is key to optimizing outcomes. By taking a cost and benefit perspective to the use of GenAI, we extend research focused on the potential productivity benefits of GenAI. We identified the psychological cost that workers risk incurring when leveraging GenAI in different collaboration roles. By accounting for the psychological consequences of using GenAI, we extend the literature to account for multiple goals held by both organizations and employees.

Our findings suggest that the critical question is not whether GenAI enhances productivity (it does), but at what cost? We found that the cost incurred depended on the specific GenAI collaboration strategy

employed. Thus, we extend work examining different GenAI modalities (Chen and Chan 2024) by 1) comparing GenAI collaboration strategies to each other (rather than comparing them each to the absence of GenAI), 2) identifying the psychological costs associated with each, 3) identifying the optimal strategy, and 4) examining whether people select the optimal strategy. Moreover, we focus on the GenAI collaboration strategies fundamental to writing: drafting and editing.

We further examined the mechanics of GenAI-human collaborations, specifically, how each collaborator spent their time. Past work comparing the presence and absence of the tool showed that GenAI improved productivity by both improving quality and saving time. In contrast, we found that using GenAI improved quality but did not save time. We replicated the benefit to productivity when introducing GenAI to the workflow and further discovered that productivity did not depend on the GenAI collaboration strategy employed. Within our experiments, both strategies were similarly productive.

More broadly, when new technology is introduced, research seeks to understand how the presence of the tool changes our lives and experiences. But as technology proliferates, it is important to understand the different *ways* that people utilize it, especially if benefits are associated with a cost. For example, work on how technology affects people's happiness first examined the presence versus absence of smartphones during social interactions (Misra et al. 2016, Dwyer et al. 2018). More recently, work has examined *how* different ways of utilizing smartphones influence users (Kushlev and Leita0 2020, Leita0 et al. 2024) and *how* different ways of using social media blocking applications can backfire (Chen and Chan 2026). In the next section, we extend our contributions by broadening our view to additional GenAI collaboration strategies. Then, we discuss organizational implications, followed by limitations and future directions.

Broader Set of Strategies for GenAI Collaboration. In order to further inform prescriptions for utilizing GenAI, we examined two additional strategies: *retrospective GenAI editor* and *GenAI only*. This helped identify which strategies best align with individual and organizational goals from a broader set of options. Specifically, we tested 1) can GenAI improve people's work without the worker's knowledge or involvement, post-hoc? and 2) how does completely automating the task affect productivity compared to the different strategies?

First, we tested a common use case: managers enhancing employee deliverables via GenAI or teachers adapting colleagues' plans through GenAI. We input lesson plans from our control condition to GenAI for editing. Thus, we created a new GenAI collaboration strategy: the retrospective GenAI editor strategy. We hypothesized that retrospective editing would produce higher-quality deliverables because people would potentially invest greater effort in initial drafts when they did not expect GenAI assistance (our control condition). We predicted that this strategy would produce the best of both human and GenAI contributions by leveraging GenAI to enhance a deliverable crafted via maximum human effort.

This hypothesis was not supported. While people spent marginally more time in the control condition than in the GenAI editor condition to produce their lesson plans, the additional time spent did not result

in higher quality on the final plan after GenAI edited the work. Judges scored teaching plans from the retrospective GenAI editor strategy similarly to those generated by using GenAI as an editor (average judge score difference = $5.95e^{-03}$, $t(365.34) = .08$, $p = .940$; Cohen's $d = 7.88e^{-03}$, $BF_{01} = 9.022$). Surprisingly, there were no differences in average text length between GenAI editor and retrospective GenAI drafts (word difference = 2.48, $t(343.26) = .29$, $p = .773$; Cohen's $d = .03$, $BF_{01} = 8.70$). These findings suggest that GenAI adapted its editing relatively uniformly regardless of the initial draft quality that it received as input. The retrospective strategy combines the time investment of creating unaided work with the time needed for GenAI to edit, yielding lower productivity than direct collaboration, without any quality advantages. While intuitively appealing as a way to enhance existing work, this strategy offered no advantages over direct collaboration and represented an inefficient hybrid strategy.

Second, in order to examine a GenAI only strategy, we revisited the GenAI drafter condition and analyzed the draft produced by GenAI. The GenAI only strategy produced higher quality plans compared to using GenAI as a drafter (difference = $-.32$, $t(284) = -3.57$, $p < .01$, Cohen's $d = -.21$; $BF_{01} = 1.16e^{-05}$). Quality decreased when people edited the GenAI draft and GenAI took only 41 seconds to generate the plan (versus 6.6 minutes total when using GenAI to draft) ⁶. Leveraging GenAI on its own, without human collaboration, maximized immediate productivity.

Despite this productivity advantage, the GenAI Only strategy has significant limitations. Even if a task is automated, organizations must balance data and human judgment, which requires people in the workflow. Organizations need people to: 1) judge whether input data is relevant to organizational goals and problems, 2) check that outputs are aligned with organizational goals and thus action-worthy, 3) decide whether, when, and in what context to deploy it, and in many cases, 4) execute on insights and plans produced by GenAI (Logg 2022).

While we could not directly measure the psychological experience of work for deliverables created solely with GenAI, we predict that using the GenAI only strategy would yield even lower control, responsibility, and creativity than the GenAI drafter condition. Furthermore, the disconnect between the output in the earlier stages of the workflow and the person acting on that output can create additional difficulties. For example, teachers who must deliver these lesson plans in their classrooms might experience a lack of ownership, which could undermine effective implementation and exacerbate disengagement. Teachers unfamiliar with the content and unprepared to adapt plans to student needs may struggle during instruction, potentially offsetting any time saved from the creation phase.

Organizational Implications. Many organizations have invested massive amounts of resources in order to provide employees with GenAI tools, expecting that such investments will directly translate into organizational success (Dutt et al. 2026). However, simply providing GenAI tools without guidance about how to

⁶ As noted in Table 9, participants who had GenAI edit reviewed the GenAI output in order to forecast the average judge score.

leverage them is insufficient. Whether organizations can effectively leverage the power of GenAI depends on how employees use GenAI. Yet, many managers ignore the meaningful psychological costs incurred when adding GenAI to the workflow. If organizations continue to ignore *how* people use GenAI, they risk employees selecting suboptimal strategies.

Data analytics needs psychology (Logg 2022) and so too does GenAI. If organizations do not understand how people collaborate with GenAI and how GenAI affects employees' psychology, they cannot realize the full potential of GenAI. Theory papers and special issues call for leveraging GenAI beyond just productivity goals (Von Krogh 2018, Barnes et al. 2022). In fact, there are broader calls for incorporating psychology and behavioral science into the creation of algorithms and GenAI (Lu et al. 2019, Guszczka et al. 2022, Sacher et al. 2026), including auditing algorithms (Logg 2019, 2022, Hauser et al. 2025). Some work tests the impact on creativity for users who collaborate with GenAI in creative work and the contributions made to the broader culture (Doshi and Hauser 2024, Zhou and Lee 2024, Zhou et al. 2025), how authentically third parties view work when creators collaborate with algorithms (Jago 2019, Jago et al. 2022), and how much credit people receive in GenAI-human collaborations (Jago and Carroll 2024).

Understanding the unintended consequences of GenAI, especially psychological ones, is important. A theoretical review of research from psychology as well as human-computer interaction calls for research on how interacting with GenAI influences people's agency, autonomy, authenticity, and empathy (Pentina et al. 2023). Another theoretical paper outlines how algorithms provide organizations control and could potentially influence feelings of disempowerment, discrimination, and stress (Kellogg et al. 2020). Some work answers these calls by showing how using algorithms as mediators between workers and managers influences workers' intrinsic motivation (Chan et al. 2023) and how managers and workers view roles which are algorithmically managed (Jago et al. 2024, Schweitzer and De Cremer 2024). We contribute to work addressing these calls by examining how GenAI collaboration strategies affect people's psychological experience of work.

Our results have clear organizational implications: including GenAI in the workflow damages employees' feelings of control and responsibility. The results also provide a prescription: when employees use GenAI, as they already are, the optimal strategy is to use GenAI to edit. Using GenAI to edit (rather than draft) allows people to preserve their psychological experience of work. However, people select the suboptimal strategy, using GenAI to draft. This suggests that organizations should provide employees guidance on their GenAI collaboration strategies and do so while considering organizational and project goals. Lastly, at times when organizations consider productivity their primary goal, our results suggest that either strategy suffices.

Limitations and Future Directions. Our work indicates multiple directions of future inquiry: How does GenAI change how we see ourselves? What drives the selection of different GenAI collaboration strategies? How does the type of task influence which strategy is optimal? What other psychological experiences of work are influenced by strategy? When is GenAI use beneficial to the psychological experience?

Effect of GenAI on Self-Perceptions. How does GenAI change how we see ourselves? Future research should further examine how GenAI influences people, specifically their self-perceptions. We predict that using GenAI changes how people view their own competency at a task and their broader topic expertise. It could also influence people's self-perceptions for social interactions. Does using GenAI influence how people view how cooperative they are when collaborating with other people? Understanding how GenAI influences social interactions with other people can further help organizations identify the optimal GenAI collaboration strategy in situations when they need to consider goals that are relevant beyond just the current workflow.

Relatedly, we predict that GenAI is a promising motivational tool to solve the “starting problem.” For example, when staring at a blank document for an important task, one might experience writer's block. While people tend to put off work to progress towards goals (O'Donoghue and Rabin 1999), research shows that using commitment devices (Milkman et al. 2008) and using events as “fresh starts” (Dai et al. 2014) can help people make beneficial choices between what they “want” to do and what they “should” do (Bazerman et al. 1998, Milkman et al. 2009, 2010). Can people use GenAI as an intervention to help start a task? We predict that interacting with GenAI helps people feel better prepared to start a task and may help people stop procrastinating tasks.

Future work should test the extent to which GenAI can help people “jumpstart” tasks and develop ideas by lowering their threshold for developing an idea. Simply sharing an idea with GenAI via short prompts might serve as a motivator in and of itself, perhaps without even using its response. Further, work should test how using GenAI to increase motivation to action differs from “bouncing ideas” off of other people, comparing GenAI and humans in the efficacy of helping people to start looming tasks.

Mechanisms Driving GenAI Collaboration Strategy Selection. Understanding why workers gravitated towards a strategy that diminished their psychological experience of work represents a key direction for future research, as organizations may need to actively guide employees toward collaboration strategies rather than assume that they will select the optimal strategy on their own. Do people underestimate just how much collaborating with GenAI diminishes their psychological experience of work? Further, how resilient are people to these psychological costs over time? How can people buffer themselves? Perhaps people can buffer themselves by focusing their time on other fulfilling tasks, decisions, and roles in order to help maintain engagement and resilience. Future work should test people's broader expectations for different GenAI collaboration strategies. For instance, their preference to have GenAI draft may reflect several additional factors: lower perceived difficulty in editing text from GenAI relative to creating an original draft, perceptions of GenAI errors that lead them to prefer to “check” its work, or simply exploring the capabilities of a new technology.

Types of Tasks and Collaboration Strategies. We should note that our experiments examined a one-shot collaboration with GenAI. We employed this simplified setting in order to standardize the collaboration and

create conditions that were more comparable. Future research should examine longer time horizons, different modes of collaboration (Dhillon et al. 2024), as well as the specific stages within multi-round collaborations. For example, do people who first use GenAI to draft iterate through more of revisions with the tool than when they start with GenAI editing their work? Relatedly, work should examine which strategies, and which sequence of strategies, are most sustainable over time, similar to how (Anwar et al. 2025) examined long-term satisfaction when using recommendation systems that are optimized for short-term engagement and how (Yang et al. 2024) examined potential long-term policy consequences. Although we began to evaluate user frustration, future work should test which strategies frustrate people enough that they either switch strategies or stop collaborating with GenAI altogether. Instead, people may seek out other humans with whom to work or to simply work on their own.

Additionally, our experiments focused on moderately complex creative tasks. Depending on the sector, employees may integrate GenAI differently, from judges drafting opinions to consultants producing client deliverables to researchers building literature reviews. The nature of work (routine versus non-routine) may affect GenAI use (Brown et al. 2025). The extent to which our findings generalize across task types, expertise levels, and professional contexts remains an open question. For instance, daily work tasks vary in creativity requirements, verifiability of outputs, and implementation demands. Future work should test whether these aspects, along with organizational goals, may produce different trade-offs (Wang et al. 2023) between collaboration strategies.

Influence on Other Aspects of the Psychological Experience of Work. We measured control, responsibility, and creativity to assess the psychological experience of work, but there are additional psychological dimensions related to the experience of work to consider. There are two broad dimensions that the future work should evaluate: *cognitive impact*, detailing how GenAI impacts the function of tasks, and *social impact*, detailing how individuals feel in relation to their work. Research into the cognitive impact of GenAI should evaluate changes such as cognitive load, depletion, and learning. Research into the social impact of GenAI should evaluate autonomy, ownership (Gero et al. 2026), and meaning/purpose. Each of these aspects (and others) may be directly affected by this new medium of work, with potential gains and losses from using GenAI.

4. Conclusion

Our findings are crucial for creating organizational environments that encourage GenAI use that maximizes employees' productivity and preserves their psychological experience of work. The discussions around GenAI deployment in organizations have focused on providing access to employees. However, because managers are overlooking how employees actually collaborate with GenAI, they are potentially unaware of the significant costs to their employees' psychological experience of work. We find that collaboration between GenAI and people is not straightforward, but a balance between productivity benefits and psychological

harms. Although using GenAI benefited productivity regardless of strategy, leveraging it was detrimental to people's control and responsibility. People further harmed their experience of responsibility and creativity when using GenAI to draft. Through our work, we identified the optimal GenAI collaboration strategy: using GenAI to edit. Our results suggest that contingent on using GenAI, people can avoid a further cost to their psychology by using GenAI to edit rather than draft, a strategy that people did not choose for themselves.

References

- Alfaro E, Cabrales A, Durán Roa JE, Garicano L, Pérez del Caño I, Roldán Monés T, Vieira de Santiago G (2026) The hidden demand for AI inside your company. *Harvard Business Review* URL <https://store.hbr.org/product/the-hidden-demand-for-ai-inside-your-company/H095G0>, online article.
- Amabile TM, Pratt MG (2016) The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior* 36:157–183.
- Anwar MS, Dhillon PS, Schoenebeck G (2025) Recommendation and temptation. *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 422–431.
- Barnes SJ, Guo Y, Chan J (2022) Big data analytics for sustainability: Insight through technological innovation. *Information & Management* 59(5):103627.
- Bazerman MH, Tenbrunsel AE, Wade-Benzoni K (1998) Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review* 23(2):225–241.
- Brecheisen J (2024) What companies don't know about how workers use AI. *Harvard Business Review* URL <https://hbr.org/2024/05/research-what-companies-dont-know-about-how-workers-use-ai>, online article.
- Brown AS, Dishop CR, Kuznetsov A, Chao PY, Woolley AW (2025) Beyond efficiency: Trust, AI, and surprise in knowledge work environments. *Computers in Human Behavior* 167:108605.
- Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *The Quarterly Journal of Economics* 140(2):889–942.
- Chan J, Kyung N, Yoon S, Kim Y (2023) Algorithm as boss or coworker? Randomized field experiment on algorithmic control and collaboration in gig platform. *ICIS 2023 Proceedings*, number 15.
- Chatterji A, Cunningham T, Deming DJ, Hitzig Z, Ong C, Shan CY, Wadman K (2025) How people use ChatGPT. Technical report, National Bureau of Economic Research.
- Chen Z, Chan J (2024) Large language models in creative work: The role of collaboration modality and user expertise. *Management Science* 70(12):9101–9117.
- Chen Z, Chan J (2026) Less is not always more: The impact of social media block intensity and immediacy on work time. *Information Systems Research* Ahead of Print.
- Chew WB (1988) No-nonsense guide to measuring productivity. *Harvard Business Review* 66:110–118.
- Cu MA, Hochman S (2023) Scores of Stanford students used ChatGPT on final exams, survey suggests. *Stanford Daily*. <https://stanforddaily.com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests>.
- Dai H, Milkman KL, Riis J (2014) The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science* 60(10):2563–2582.
- Danna K, Griffin RW (1999) Health and well-being in the workplace: A review and synthesis of the literature. *Journal of Management* 25(3):357–384.

- Dell'Acqua F, McFowland III E, Mollick E, Lifshitz H, Kellogg KC, Rajendran S, Kraye L, Candelon F, Lakhani KR (2026) Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science* 37(2):403–423.
- Den Houter K (2024) AI in the workplace: Answering 3 big questions. Technical report, Gallup, URL <https://www.gallup.com/workplace/651203/workplace-answering-big-questions.aspx>.
- Dhillon PS, Molaei S, Li J, Golub M, Zheng S, Robert LP (2024) Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.
- Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10(28):eadn5290.
- Dutt A, Rapoport G, Chatterji A, Weeratunga G, Satcher H (2026) How to move from AI experimentation to AI transformation. *Harvard Business Review* URL <https://hbsp.harvard.edu/product/H0962R-PDF-ENG>.
- Dwyer RJ, Kushlev K, Dunn EW (2018) Smartphone use undermines enjoyment of face-to-face social interactions. *Journal of Experimental Social Psychology* 78:233–239.
- Flower L, Hayes JR (1981) A cognitive process theory of writing. *College Composition & Communication* 32(4):365–387.
- Gallup (2025) State of the global workplace: 2025 report. Technical report, Gallup, Inc., URL <https://www.gallup.com/file/workplace/659528/state-of-the-global-workplace-2025-download.pdf>.
- Gallup (2026) State of the global workplace: 2026 report. Technical report, Gallup, Inc., URL <https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx>.
- Gero KI, Long T, Schnitzler C, Dhillon P (2026) From planning to revision: How AI writing support at different stages alters ownership. *arXiv preprint arXiv:2604.11009*.
- Gong Y, Huang JC, Farh JL (2009) Employee learning orientation, transformational leadership, and employee creativity: The mediating role of employee creative self-efficacy. *Academy of Management Journal* 52(4):765–778.
- Gonzalez C, Admoni H, Brown S, Woolley AW (2025) COHUMAIN: Building the socio-cognitive architecture of collective human–machine intelligence.
- Gonzalez C, Donahue K, Goldstein DG, Heidari H, Jalali MS, Schelble B, Singh A, Woolley AW (2026) Toward a science of human–AI teaming for decision making: A complementarity framework. *PNAS nexus* 5(3):pgag030.
- Grant AM, Christianson MK, Price RH (2007) Happiness, health, or relationships? Managerial practices and employee well-being tradeoffs. *Academy of Management Perspectives* 21(3):51–63.
- Guszcza J, Danks D, Fox CR, Hammond KJ, Ho DE, Imas A, Landay J, Levi M, Logg J, Picard RW, et al. (2022) Hybrid intelligence: A paradigm for more responsible practice. *Available at SSRN 4301478*.
- Hackman JR, Oldham GR (1976) Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* 16(2):250–279.

- Hauser OP, Light M, Shelmerdine L, Blumenau J (2025) Why evaluating the impact of AI needs to start now. *Nature* 643(8073):910–912.
- Humphrey SE, Nahrgang JD, Morgeson FP (2007) Integrating motivational, social, and contextual work design features: A meta-analytic summary and theoretical extension of the work design literature. *Journal of Applied Psychology* 92(5):1332.
- Indeed (2026) Customer service job postings on indeed in the united states. Retrieved from FRED, Federal Reserve Bank of St. Louis, accessed May 13, 2026. <https://fred.stlouisfed.org/series/IHLIDXUSTPCUSTSERV>.
- Jago AS (2019) Algorithms and authenticity. *Academy of Management Discoveries* 5(1):38–56.
- Jago AS, Carroll GR (2024) Who made this? Algorithms and authorship credit. *Personality and Social Psychology Bulletin* 50(5):793–806.
- Jago AS, Carroll GR, Lin M (2022) Generating authenticity in automated work. *Journal of Experimental Psychology: Applied* 28(1):52.
- Jago AS, Raveendhran R, Fast N, Gratch J (2024) Algorithmic management diminishes status: An unintended consequence of using machines to perform social roles. *Journal of Experimental Social Psychology* 110:104553.
- Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: The new contested terrain of control. *Academy of Management Annals* 14(1):366–410.
- Kushlev K, Leitao MR (2020) The effects of smartphones on well-being: Theoretical integration and research agenda. *Current Opinion in Psychology* 36:77–82.
- Leitao MR, Proulx JD, Kushlev K (2024) Smartphones undermine social connectedness more in men than women: A mini mega-analysis. *Technology, Mind, and Behavior* 5(1):15.
- Logg JM (2019) Using algorithms to understand the biases in your organization. *Harvard business review* 1–2.
- Logg JM (2022) The psychology of big data: Developing a “Theory of Machine” to examine perceptions of algorithms. Matz SC, ed., *The Psychology of Technology: Social Science Research in the Age of Big Data*, 349–378 (American Psychological Association).
- Logg JM, Dorison CA (2021) Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes* 167:18–27.
- Lu J, Lee D, Kim TW, Danks D (2019) Good explanation for algorithmic transparency. Available at SSRN 3503603 .
- Luca M, Kleinberg J, Mullainathan S (2016) Algorithms need managers, too. *Harvard Business Review* 94(1):20.
- Martin R (2015) What economists get wrong about measuring productivity. *Harvard Business Review* URL <https://store.hbr.org/product/what-economists-get-wrong-about-measuring-productivity/H02C0X>.
- Mazeika M, Gatti A, Menghini C, Schwag UM, Singhal S, Orlovskiy Y, Basart S, Sharma M, Peskoff D, Lau E, Lim J, Carroll L, Blair A, Sivakumar V, Basu S, Kenstler B, Ma Y, Michael J, Li X, Ingebretsen O, Mehta A, Mottola J, Teichmann J, Yu K, Shaik Z, Khoja A, Ren R, Hausenloy J, Phan L, Htet Y, Aich A, Rabbani T, Shah V,

- Novykov A, Binder F, Chugunov K, Ramirez L, Geralnik M, Mesura H, Lee D, Cardona EYH, Diamond A, Yue S, Wang A, Liu B, Hernandez E, Hendrycks D (2025) Remote labor index: Measuring AI automation of remote work. URL <https://arxiv.org/abs/2510.26787>.
- Milkman KL, Rogers T, Bazerman MH (2008) Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science* 3(4):324–338.
- Milkman KL, Rogers T, Bazerman MH (2009) Highbrow films gather dust: Time-inconsistent preferences and online DVD rentals. *Management Science* 55(6):1047–1059.
- Milkman KL, Rogers T, Bazerman MH (2010) I'll have the ice cream soon and the vegetables later: A study of online grocery purchases and order lead time. *Marketing Letters* 21(1):17–35.
- Misra S, Cheng L, Genevie J, Yuan M (2016) The iPhone effect: The quality of in-person social interactions in the presence of mobile devices. *Environment and Behavior* 48(2):275–298.
- Morey RD, Rouder JN (2024) *BayesFactor: Computation of Bayes Factors for Common Designs*. URL <https://CRAN.R-project.org/package=BayesFactor>, R package version 0.9.12-4.7.
- Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.
- O'Donoghue T, Rabin M (1999) Doing it now or later. *American Economic Review* 89(1):103–124.
- Pentina I, Xie T, Hancock T, Bailey A (2023) Consumer-machine relationships in the age of artificial intelligence: Systematic literature review and research directions. *Psychology & Marketing* 40(8):1593–1614.
- Qian C, Wexler J (2024) Take it, leave it, or fix it: Measuring productivity and trust in human-AI collaboration. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 370–384.
- R Core Team (2025) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>.
- Revelle W (2026) *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, URL <https://CRAN.R-project.org/package=psych>, R package version 2.6.1.
- Sacher PM, Michie S, Hauser OP, Ferrère A, Salzer S, Rodger A, Borg JS, Pogrebna G, Murphy SA (2026) The missing discipline in AI: a call for behavioural science. *Wellcome Open Research* 11:152.
- Schweitzer S, De Cremer D (2024) When being managed by technology: does algorithmic management affect perceptions of workers' creative capacities? *Academy of Management Discoveries* 10(3):375–392.
- Simmons JP, Nelson LD, Simonsohn U (2012) A 21 word solution. Available at SSRN 2160588 .
- Tamkin A, McCrory P (2025) Estimating AI productivity gains from Claude conversations. URL <https://www.anthropic.com/research/estimating-productivity-gains>.
- Von Krogh G (2018) Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries* 4(4):404–409.

- Wang Q, Huang Y, Jasin S, Singh PV (2023) Algorithmic transparency with strategic users. *Management Science* 69(4):2297–2317.
- World Economic Forum (2025) Future of jobs report 2025. Technical report, World Economic Forum, Geneva, Switzerland, URL <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>.
- Wu S, Liu Y, Ruan M, Chen S, Xie XY (2025) Human-generative AI collaboration enhances task performance but undermines human’s intrinsic motivation. *Scientific Reports* 15(1):15105.
- Yang J, Eckles D, Dhillon P, Aral S (2024) Targeting for long-term outcomes. *Management Science* 70(6):3841–3855.
- Zao-Sanders M (2025) How people are really using Gen AI in 2025. *Harvard Business Review* URL <https://hbsp.harvard.edu/product/H080PN-PDF-ENG>, online article.
- Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art. *PNAS nexus* 3(3):pgae052.
- Zhou EB, Lee D, Gu B (2025) Who expands the human creative frontier with generative AI: Hive minds or masterminds? *Science Advances* 11(36):eadu5800.

Appendix A: Study 1

A.1. t-test & Bayesian Analysis

A.1.1. Productivity Comparing levels of Productivity.

Table 11 Independent Samples t-test Comparing the Productivity Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	0.11	-4.81	-4.91	2.09	0.04	598.12	[.01; .21]	.17	1.35
Expertise	Low-High	0.07	-4.83	-4.89	1.25	.21	598.93	[-.04; .17]	.10	5.12

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.1.2. Avg. Judge's Score Comparing Average Judge's Score.

Table 12 Independent Samples t-test Comparing the Judge's Score Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	-0.08	5.57	5.65	-1.24	.22	584.23	[-.21; .05]	-.10	5.25
Expertise	Low-High	-0.01	5.61	5.61	-.14	.89	591.46	[-.14; .12]	-.01	10.90

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.1.3. Time On Survey Comparing Time on Survey.

Table 13 Independent Samples t-test Comparing the Time Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	-.13	6.51	6.64	-2.69	0.01	598.81	[-.23; -.04]	-.22	.33
Expertise	Low-High	-.07	6.53	6.61	-1.46	.14	598.98	[-.17; .02]	-.12	3.87

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.1.4. Control Comparisons of Feelings of Control in the process.**Table 14** Independent Samples t-test Comparing the Control Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	-9.88	59.53	69.41	-4.62	<.01	579.26	[-14.08; -5.68]	-.38	.00046
Expertise	Low-High	-5.37	61.61	66.98	-2.46	.01	591.18	[-9.65; -1.08]	-.20	.58

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.1.5. Responsibility Comparisons of Feelings of Responsibility toward outcome.**Table 15** Independent Samples t-test Comparing the Responsibility Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	-11.04	51.59	62.63	-5.07	<.01	584.93	[-15.32; -6.77]	-.42	$5.81e^{-5}$
Expertise	Low-High	-.07	56.88	56.95	-.03	.97	598.81	[-4.47; 4.32]	$2.66e^{-3}$	11

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.1.6. Creativity Comparisons of Creativity utilized.**Table 16** Independent Samples t-test Comparing the Productivity Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role	Drafter-Editor	-20.48	49.16	69.64	-9.60	<.01	558.64	[-24.67; -16.29]	-.81	$1.38e^{-17}$
Expertise	Low-High	-.39	58.84	59.24	-0.17	.87	595.21	[-4.94; 4.15]	-.01	10.82

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.2. Miscalibration of Change

Participant's ability to judge how their score changes between the initial and final draft

A.2.1. Productivity Comparing levels of Productivity.

Table 17 Independent Samples t-test Comparing the Productivity Between Conditions

	Comparison	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GPT Role Expertise	Drafter-Editor	.14	.53	.38	.91	.36	593.60	[−.17; .45]	.08	7.31
	Low-High	-.06	.43	.49	-.36	.72	598.64	[−.37; .25]	-.03	10.31

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

A.3. Robustness Checks

A.3.1. Those with the only middle 95% of the number of words composed The two tables below represent the analysis of our main dependent variables, excluding those in the highest and lowers percentile of words submitted.

Table 18 OLS Regression of Participant Score and Productivity Measures

	<i>Dependent variable:</i>		
	Log(Productivity)	Avg. Judge's Score	Log(Time)
	(1)	(2)	(3)
Intercept	−4.50*** (.14)	5.93*** (.16)	6.28*** (.14)
Expertise - High	−.07 (.05)	−.02 (.06)	.07 (.05)
GenAI Role - Drafter	.07 (.05)	−.05 (.06)	−.09 (.05)
Comfort w/ Writing	−.003* (.001)	−.001 (.001)	.003* (.001)
Task Difficulty	−.004*** (.001)	−.002* (.001)	.004*** (.001)
Exprnce_1_5Extensiv	.01 (.03)	−.01 (.03)	−.01 (.02)
Observations	574	574	574
R ²	.04	.01	.04
Adjusted R ²	.03	−.001	.03
Residual Std. Error (df = 568)	.63	.72	.60
F Statistic (df = 5; 568)	4.53***	.92	4.39***

Note: *p<0.05; **p<0.01; ***p<0.001

Table 19 OLS Regression of Participant Score and Productivity Measures

	<i>Dependent variable:</i>		
	Control (1)	Responsibility (2)	Creativity (3)
Intercept	28.38*** (5.67)	8.85 (5.67)	16.17** (5.49)
Expertise - High	4.88* (2.07)	-.09 (2.07)	-.18 (2.01)
GenAI Role - Drafter	-7.07*** (2.10)	-7.47*** (2.11)	-16.59*** (2.04)
Comfort w/ Writing	.39*** (.05)	.43*** (.05)	.51*** (.05)
Task Difficulty	.09* (.04)	.25*** (.04)	.24*** (.04)
Experience w/ AI	1.77 (1.03)	3.32** (1.03)	1.78 (1.00)
Observations	574	574	574
R ²	.14	.19	.28
Adjusted R ²	.14	.18	.28
Residual Std. Error (df = 568)	24.82	24.85	24.03
F Statistic (df = 5; 568)	19.05***	26.14***	44.52***

Note: *p<0.05; **p<0.01; ***p<0.001

Appendix B: Study 2

B.1. t-tests & Bayesian Analysis

B.1.1. Productivity Comparing the level of Productivity.

Table 20 Independent Samples t-test Comparing the Productivity Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-0.05	-3.92	-3.87	-0.65	0.52	377.50	[-.21; .11]	-.23	7.71
Control-Editor	-0.27	-4.14	-3.87	-3.13	<.01	385.43	[-.44; -.10]	-.32	.09
Control-Drafter	-0.22	-4.14	-3.92	-2.82	<.01	485.98	[-0.37; -0.07]	-.26	.21

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.1.2. Avg. Judge's Score Comparing the of Average Judge's Score, both within and between conditions.

Comparing the differences between the initial and final drafts

Table 21 Paired Samples t-test Comparing the Judge's Score in the Initial and Final Drafts

	Avg. Δ	t	p value	df	95% Confidence	Cohen's D	BF_{01}
GenAI Editor	.78	6.99	<.01	177	[.56; 1.00]	.53	$8.27e^{-38}$
GenAI Drafter	-.32	-3.57	<.01	284	[-.50; -.14]	-.21	$1.16e^{-05}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

Table 22 Independent Samples t-test Comparing the Judge's Score Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-.23	5.86	6.09	-2.46	.01	458.09	[-.42; -.05]	-.23	.809
Control-Editor	-.67	5.42	6.09	-7.10	<.01	405.43	[-.86; -.49]	-.70	$3.90e^{-9}$
Control-Drafter	-.44	5.42	5.86	-4.33	<.01	505.86	[-.64; -.24]	-.39	$1.52e^{-3}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

Table 23 Independent Samples t-test Comparing the Judge's Score Directly After AI Intervention

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	0.09	6.18	6.09	1.08	0.28	404.35	[-0.07; 0.25]	.11	5.53

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.1.3. Time On Task Comparing Time on Task.

Table 24 Independent Samples t-test Comparing the Time Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-.02	5.65	5.67	-.21	.83	366.85	[-.17; .14]	-.02	9.26
Control-Editor	.13	5.80	5.67	1.55	.12	378.54	[-.03; .29]	-.70	2.83
Control-Drafter	0.15	5.80	5.65	2.02	.04	489.76	[.00; .29]	.18	1.41

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.1.4. Control Comparisons based on Feelings of Control over the process.**Table 25** Independent Samples t-test Comparing the Feelings of Control Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-1.23	74.67	75.89	-.70	.48	429.65	[-4.67; 2.21]	-.07	7.61
Control-Editor	10.11	86.00	75.89	6.51	<.01	337.87	[7.06; 13.17]	.71	$9.96e^{-9}$
Control-Drafter	11.34	86.00	74.67	7.51	<.01	501.65	[8.37; 14.30]	.67	$3.13e^{-10}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.1.5. Responsibility Comparisons based on Feelings of Responsibility toward outcome.**Table 26** Independent Samples t-test Comparing the Feelings of Responsibility Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-7.56	70.91	78.47	-3.91	<.01	443.13	[-11.36; -3.76]	-.37	.01
Control-Editor	9.72	88.20	78.47	5.91	<.01	329.37	[6.49; 12.96]	.65	$2.81e^{-7}$
Control-Drafter	17.29	88.20	70.91	10.37	<.01	483.20	[14.01; 20.56]	.94	$1.06e^{-18}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.1.6. Creativity Comparisons based on Feelings of Creativity utilized.**Table 27** Independent Samples t-test Comparing the Feelings of Creativity Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	-16.65	58.39	75.04	-7.89	<.01	436.50	[-20.80; -12.50]	-.75	$5.54e^{-11}$
Control-Editor	2.97	78.01	75.04	1.61	0.11	337.83	[-0.66; 6.61]	.18	2.43
Control-Drafter	119.63	78.01	58.39	10.72	0.00	496.54	[16.03; 23.22]	.96	$4.06e^{-20}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.2. Mediation Analysis

We conducted a mediation analysis of the Number of words on the effect of AI Used predicting Avg. Judge's Score

Table 28 Exp. 2: Mediating Effect of # Words on AI Used for Avg. Judge's Score

	Estimate	95% Confidence	p value
Average Causal Mediation Effect (ACME)	.57	[.45;.68]	< .001 ***
Average Direct Effect (ADE) - AI Used (Post-Mediated)	-0.07	[-.21;.14]	.726
Total Effect of AI Used (Pre-Mediation)	0.53	[.37;.70]	< .001 ***
Proportion Mediated	1.06	[.78;1.54]	< .001 ***

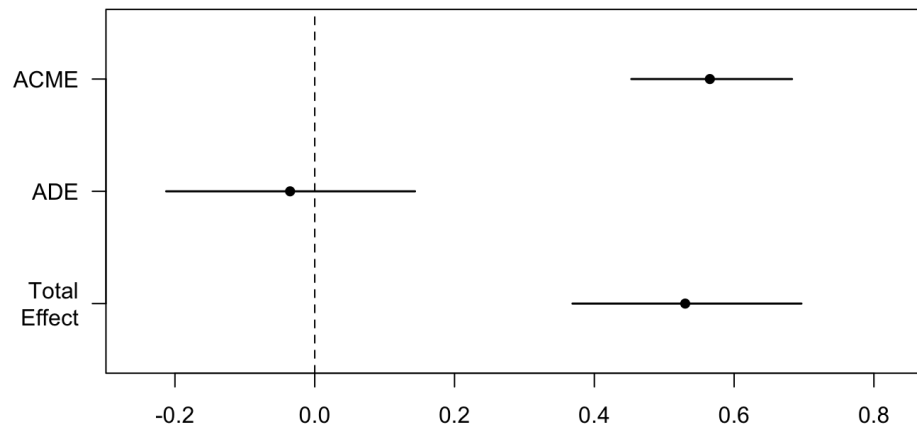


Figure 5 Plot of Mediated Effect of # Words on AI used for Avg. Judge's Score

B.3. Write Time

Comparing only writing time between the Control and GenAI Editor conditions.

Table 29 Independent Samples t-test Comparing the Write Time Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Control-Editor	0.41	5.80	5.39	4.38	<.01	335.39	[.23; .59]	.48	$6.23e^{-4}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.
Write time is log transformed.

B.4. Frustration

Tables illustrating people's feelings of Motivation after creating the lesson plans.

Table 30 Independent Samples t-test Comparing the Feelings of Frustration with the Process Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	1.01	20.37	19.36	0.44	0.66	344.35	[-3.49; 5.51]	.05	8.28
Control-Editor	5.24	24.60	19.36	2.15	.03	363.42	[.45; 10.02]	.23	1.03
Control-Drafter	4.23	24.60	20.37	1.94	.05	475.03	[-.05; 8.50]	.18	1.58

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.5. Number of Words - Retrospective GenAI/GenAI Editor

Table 31 Independent Samples t-test Comparing the Number of Words Between Retrospective GenAI and GenAI Editor

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Retro Control-Editor	2.48	451.18	448.69	.29	0.77	343.26	[-14.44; 19.41]	.03	8.70

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.6. Decomposing Judge's Score

B.6.1. GenAI Editor Change in Judges Score by Criteria for GenAI Editor Condition only

Table 32 Paired Samples t-test Comparing the Decomposed Judge's Score in the Final versus Initial drafts

	Avg. Δ	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Avg. Judge Score	.78	6.99	<.001	177	[.56; 1.00]	.53	$7.08e^{-9}$
Clear Objectives	10.74	6.81	<.001	177	[7.63; 13.85]	.51	$1.81e^{-8}$
Content Quality	12.65	7.54	<.001	177	[9.34; 15.95]	.57	$3.25e^{-10}$
Feasibility	5.89	3.93	<.001	177	[2.93; 8.85]	.30	$8.27e^{-3}$
Student Appeal	9.53	5.79	<.001	177	[9.53; 12.78]	.44	$3.19e^{-6}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.6.2. Retrospective GenAI Change in Judges Score by Criteria for Retrospective GenAI Condition only

Table 33 Paired Samples t-test Comparing the Decomposed Judge's Score in the Final versus Initial drafts

	Avg. Δ	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Avg. Judge Score	.68	8.13	<.001	229	[.51; .84]	.54	$4.48e^{-12}$
Clear Objectives	10.19	8.63	<.001	229	[7.86; 12.51]	.57	$1.90e^{-13}$
Content Quality	10.60	8.11	<.001	229	[8.02; 13.17]	.54	$5.20e^{-10}$
Feasibility	4.67	3.64	<.001	229	[2.14; 7.20]	.24	$2.36e^{-2}$
Student Appeal	8.47	6.21	<.001	229	[5.78; 11.16]	.41	$3.02e^{-7}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.7. Retrospective GenAI vs. GenAI Conditions

Comparing the outcomes from the Control condition run through the same GenAI prompt and the GenAI final versions.

Table 34 Retrospective GenAI: Independent Samples t-test Comparing Avg. Judge's Score Between Conditions

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Ret.Control-Editor	0.01	6.10	6.09	0.08	0.94	365.34	[-0.15; 0.16]	$7.88e^{-03}$	9.02
Ret.Control-Drafter	0.24	6.10	5.86	2.73	0.01	484.92	[.07; .41]	.25	.38

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

Table 35 OLS Regression Predicting Different Avg. Judge Score - Retrospective GenAI

	<i>Dependent variable:</i>		
	Avg. Judge's Score		
	(1)	(2)	(3)
Intercept	6.18*** (.08)	5.97*** (.27)	5.81*** (.28)
Known AI Use	-.02 (.10)	-.05 (.10)	.12 (.15)
AI Drafter	-.22* (.09)	-.21* (.09)	-.13 (.15)
Tenure	-.01 (.005)	-.01* (.01)	.003 (.01)
Task Comfort		.003 (.003)	.003 (.003)
Task Difficulty		-.003 (.002)	-.003 (.002)
Experience w/ AI		.03 (.04)	.03 (.04)
STEM		-.01 (.08)	-.02 (.08)
AI Used*Tenure			-.02 (.01)
AI Drafter*Tenure			-.01 (.01)
Observations	689	689	689
R ²	.02	.03	.04
Adjusted R ²	.01	.02	.03
Residual Std. Error	.98 (df = 685)	.98 (df = 681)	.98 (df = 679)
F Statistic	4.36** (df = 3; 685)	3.00** (df = 7; 681)	2.98** (df = 9; 679)

Note: *p<0.05; **p<0.01; ***p<0.001

B.8. Miscalibration of Change

Table 36 Independent Samples t-test Comparing the Predicted Score Change Versus Actual Score Change

	Difference	Est. 1	Est. 2	t	p value	df	95% Confidence	Cohen's D	BF_{01}
Drafter-Editor	1.01	20.37	19.36	0.44	0.66	344.35	[.49; 1.14]	.50	$1.22e^{-4}$

Note: BF_{01} represents likelihood for the null hypothesis over the alternative.

B.9. Robustness Checks

B.9.1. Those with the only middle 95% of the number of words composed The two tables below represent the analysis of our main dependent variables, excluding those in the highest and lowest percentile of words submitted.

Table 37 OLS Regression Predicting Productivity - Robustness Check 1

	<i>Dependent variable:</i>		
	Log(Productivity)	Avg. Judge's Score	Log(Time On Task)
	(1)	(2)	(3)
Intercept	-4.07*** (.24)	5.16*** (.27)	5.72*** (.23)
AI Used	.21* (.09)	.69*** (.10)	-.07 (.08)
GenAI Drafter	.02 (.08)	-.09 (.10)	-.05 (.08)
Tenure	-.01 (.005)	-.02** (.01)	.005 (.004)
Task Comfort	-.002 (.002)	.01 (.003)	.002 (.002)
Task Difficulty	-.002 (.001)	-.001 (.002)	.001 (.001)
Experience w/ AI	.06 (.03)	-.01 (.04)	-.07* (.03)
STEM	-.06 (.07)	.02 (.08)	.07 (.07)
Observations	653	653	658
R ²	.03	.11	.02
Adjusted R ²	.02	.10	.01
Residual Std. Error	.86 (df = 645)	.98 (df = 645)	.82 (df = 650)
F Statistic	3.18** (df = 7; 645)	10.93*** (df = 7; 645)	1.72 (df = 7; 650)

Note: *p<0.05; **p<0.01; ***p<0.001

Table 38 OLS Regression Predicting Experience - Robustness 1

	<i>Dependent variable:</i>		
	Control (1)	Responsibility (2)	Creativity (3)
Intercept	50.30*** (4.69)	49.16*** (5.21)	30.97*** (5.73)
AI Used	-10.56*** (1.72)	-9.75*** (1.92)	-3.19 (2.11)
GenAI Drafter	-1.18 (1.66)	-6.97*** (1.84)	-16.83*** (2.03)
Tenure	.05 (.09)	.08 (.10)	.14 (.11)
Task Comfort	.38*** (.05)	.40*** (.05)	.42*** (.06)
Task Difficulty	-.01 (.03)	.05 (.03)	.15*** (.04)
Experience w/ AI	1.38* (.65)	.89 (.72)	1.93* (.80)
STEM	.66 (1.35)	2.29 (1.50)	1.05 (1.66)
Observations	658	658	658
R ²	.20	.21	.24
Adjusted R ²	.19	.20	.23
Residual Std. Error (df = 650)	16.79	18.66	20.53
F Statistic (df = 7; 650)	22.57***	24.72***	29.75***

Note: *p<0.05; **p<0.01; ***p<0.001