Prochazka, Jakub; Zhou, Jing; Coita, Ioana-Florina; Akhtar, Shumi

Working Paper

A comment on "An Empirical Investigation of the Impact of ChatGPT on Creativity"

I4R Discussion Paper Series, No. 274

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Prochazka, Jakub; Zhou, Jing; Coita, Ioana-Florina; Akhtar, Shumi (2025) : A comment on "An Empirical Investigation of the Impact of ChatGPT on Creativity", I4R Discussion Paper Series, No. 274, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/333869

INSTITUTE for REPLICATION

No. 274
I4R DISCUSSION PAPER SERIES

A comment on "An Empirical Investigation of the Impact of ChatGPT on Creativity"

Jakub Prochazka
Jing Zhou
Ioanna-Florina Coita
Shumi Akhtar

December 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 274

A comment on "An Empirical Investigation of the Impact of ChatGPT on Creativity"

Jakub Prochazka[1], Jing Zhou[2], Ioanna-Florina Coita[3], Shumi Akhtar[4]

[1]Masaryk University, Brno/Czech Republic
[2]University of Edinburgh/Great Britain
[3]University of Economics in Bratislava/Slovakia
[4]The University of Sydney/Australia

DECEMBER 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A comment on

"An Empirical Investigation of the Impact of ChatGPT on Creativity"

Jakub Prochazka (Masaryk University) [1]
Jing Zhou (University of Edinburgh) [2]
Ioana-Florina Coita (University of Economics in Bratislava) [3]
Shumi Akhtar (The University of Sydney) [4]

[1] Department of Business Management, Faculty of Economics and Administration, Lipova 41a 616 00 Brno, Czech Republic, E: jak.prochazka@mail.muni.cz, corresponding author.
[2] School of Economics, University of Edinburgh, 31 Buccleuch Place, Edinburgh, EH8 9JT. E: Jing.Zhou@ed.ac.uk.
[3] Department of Finance, Faculty of Economics, University of Economics in Bratislava, Dolnozemska cesta 1, 852 35 Bratislava, Slovak Republic, E: ioanaflorina.coita@euba.sk, P: +421 267291221.
[4] Finance Discipline, The University of Sydney Business School, The University of Sydney, NSW 2006, Australia. E: shumi.akhtar@sydney.edu.au, P: +612 90369309.

Abstract

This report describes a computational reproduction of Lee and Chung's (2024) paper, which examined whether using ChatGPT (GPT-3.5) enhances creativity in adults compared to web-search assistance or no assistance. The authors presented six randomized controlled experiments showing that ChatGPT-assisted responses were assessed as significantly more creative (effect sizes ranging from Cohen's d = 0.32 to 1.88). These effects were robust across diverse tasks and contexts.

We first computationally reproduced all the main results using the original dataset and code, obtaining the same results as those presented by the authors in their paper. During the reproduction process, we identified two minor coding errors and one typographical error in the original table, none of which affected the substantive conclusions. Second, we performed a recreate reproduction for the main analysis in Experiments 1 and 3 by writing new R code. Our results again matched the results presented in the original paper.

Overall, based on our analyses, the study is fully computationally reproducible from raw data, although only with access to the original code, due to undocumented cleaning steps, some non-described exclusion criteria, and missing codebooks.

Several analyses in the original paper showed that ideas generated by ChatGPT are rated as similarly creative regardless of whether people modify them or not. We contributed to this conclusion by introducing a new robustness check using response time as a proxy for human effort in modifying ChatGPT outputs. Using data from Experiment 3, we found no significant correlation between response time and creativity in the ChatGPT condition ($r = -.079$, $p = .449$) and no moderating effect of response time on the influence of using ChatGPT on creativity. This suggests that human effort does not incrementally improve creativity beyond ChatGPT's contribution. Taken together, our findings support the original claim that using ChatGPT increases creativity regardless of the human input.

Note. The supplementary online materials for this report are available at ……

1. Introduction

The study by Lee and Chung (2024) tested the effect of using ChatGPT (GPT-3.5) on creativity of adults, using six randomized controlled experiments with creativity ratings provided by external and expert judges. The main results show that using ChatGPT assistance significantly increased creativity of ideas compared to Web-search assistance or no assistance, with effect sizes ranging from Cohen's d = 0.32 to 1.88 (see the effect sizes in Table 1).

The analyses compared how creative people are when using various technologies (ChatGPT-assisted, Web-search-assisted), or no technologies (Human-only), and how creative people are in comparison to ChatGPT. Authors understood creativity "as the generation of original and appropriate ideas (Lee & Chung, 2024, p. 1906). They asked participants to provide open answers on creativity tasks. These answers were then judged by external judges using a 7-point scale. The main Statistical analyses included t-tests, ANOVA, and mediation analysis. All analyses were done in R.

Primary data were collected using MTurk and Prolific platforms in 2023–2024 (Experiment 1: February 2023; 2A: June 2023; 2B: January 2024; 3: May 2023; 4: June 2023; 5: July 2023). Respondents were adults from the USA (mean ages ~38–43; gender roughly balanced) (p. 11).The treatment was a nudge to use ChatGPT for creative problem-solving tasks. The control conditions included the nudge to use web-search assistance (i.e., Google; in experiments 1, 2A, 3, 4, 5), no nudge (i.e., no assistance; in Experiments 2A, 2B), and also ChatGPT-generated answers (i.e., the creativity of answers processed by respondents were compared to the creativity of answers produced by ChatGPT 3.5 with no human input; in Experiment 2B). To increase the external validity of their findings, the authors also manipulated the characteristics of the creativity task in Experiments 4 and 5. In Experiment 4, the participants faced either a high- or low-constraints condition. In experiment 5, the participants were assigned to conditions that either demanded empathy or did not.

Main Claim (as stated by authors)

In the Discussion, the authors claim, “We found that using (versus not using) ChatGPT can increase the creativity of responses to problem-solving tasks. This positive effect is robust across various types of tasks, including generating creative gift ideas, repurposing items and designing innovative dining tables (p. 1910).“

This claim was supported in all 6 experiments (see Table 1) that showed that ChatGPT-assisted outputs were more creative (as evaluated by external judges) in comparison to both web-search-assisted outputs (experiments 1, 2A, 3, 4, 5) and non-assisted outputs (Experiments 2A, 2B). The effect emerged regardless of high/low constraint conditions (Experiment 4) or the need/no need for an empathic approach (Experiment 5). The authors also showed that the ChatGPT-assisted outputs were evaluated as similarly creative as ChatGPT-generated answers (Experiment 2B) in a non-preregistered supplementary analysis.

Table 1. Main results of the original study

Exp.	Condition	*M*	*SD*	*t*(df)	*p*	Cohen's *d*	*d*(95% CI)
1	ChatGPT-assisted	4.28	0.81	t(231) = 3.68	<0.001	0.49	(0.22, 0.75)
	Web-search-assisted	3.89	0.80				
2A	ChatGPT-assisted	4.56	0.75				
	Web-search-assisted	3.98	0.94	t(288) = 4.97	<0.001	0.71	(0.42, 1.00)
	Human-only	4.11	0.74	t(288) = 3.85	<0.001	0.56	(0.27, 0.85)
2B	ChatGPT-assisted	4.84	0.47	t(297) = 13.3	<0.001	1.88	(1.56, 2.20)
	ChatGPT-only	4.83	0.40	t(297) = 13.21	<0.001	1.88	(1.56, 2.20)
	Human-only	3.66	0.89				
3	ChatGPT-assisted	4.60	0.72	t(192) = 2.59	.010	0.37	(0.09, 0.66)
	Web-search-assisted	4.35	0.59				
4	ChatGPT-assisted * High constraints	4.73	0.66	t(396) = 2.25	.025	0.32	(0.04, 0.60)
	Web-search-assisted * High constraints	4.53	0.68				
	ChatGPT-assisted * Low constraints	4.62	0.57	t(396) = 2.4	.017	0.34	(0.06, 0.63)
	Web-search-assisted * Low constraints	4.40	0.61				
5	ChatGPT-assisted * High empathy	4.47	0.73	t(379) = 5.26	<0.001	0.77	(0.48, 1.07)
	Web-search-assisted * High empathy	3.83	0.97				
	ChatGPT-assisted * Baseline empathy	4.29	0.68	t(379) = 7.02	<0.001	1.01	(0.72, 1.30)
	Web-search-assisted * Baseline empathy	3.45	0.96				

Note. Reproduced from the original paper (p. 1910); the text in red is a typo, there should be "Web-searched-assisted" instead of "ChatGPT-assisted". The results of our reproduction were the same as in the original paper.

In the present report prepared as part of a collaboration between the Institute for Replication and Nature Human Behaviour (Brodeur et al., 2024), we showed that the tests supporting the above-mentioned claim are computationally reproducible using the original dataset and code, recreate reproducible using the original dataset but creating new code,

and robustness reproducible considering the real effort of the respondents measured by their response time in the creativity tasks.

For our reproduction efforts, we used the replication package cited in the original manuscript: https://osf.io/rzn87/files/osfstorage.

2. Computational and Recreate Reproducibility

The replication package contained cleaning code (complete, as part of the analysis code), analysis code (complete), and raw data (complete). There was no codebook and readme files available within the package. We were also able to obtain the cleaned data. Nevertheless, the cleaning procedure was not described in the paper or supplementary materials, so we were unable to reproduce the data cleaning without using the original codes. We found that authors used several exclusion criteria that were neither mentioned in the manuscript nor in the supplementary materials or pre-registration.

We first tried to computationally reproduce the results supporting the main claim (Table 1) in R using the original code. The reproduction was successful, as we obtained the same results as those presented in Table 1 of the original manuscript. Nevertheless, we found one typo in the original Table 1. For Experiment 4, the last row describes an interaction term “ChatGPT-assisted * Low constraints”, but the correct description of the interaction term should be “Web-search-assisted * Low constraints.” This typo did not affect the interpretation of the results or the conclusions presented in the paper. We provide the correct wording in our Table 1.

As a second step, we did a recreate reproduction by writing new code in R based on the description of the analysis in the paper and supplementary materials. For the recreate reproduction, we chose the analysis supporting the main claim in Experiments 1 and 3. The key analysis in both experiments was a Student t-test, focusing on the differences in creativity between people in ChatGPT-assisted and Web-search-assisted conditions. As there were no codebook and readme files included in the replication package, and there were no labels in the datasets, we had to inspect the original R code to identify the relevant variables (“Condition” is the independent variable; “OVERALL” is the dependent variable) within dozens of various variables. The description and the code of our

reproduction attempt are available in supplementary online materials at OSF.io. Our reproduction attempt was successful, as we drew the same conclusions as the authors and obtained the same results as the original authors (see Exp. 1 and Exp. 3 in Table 1). Generally, based on our results, we can conclude that the study is fully computationally reproducible from the raw data, but only with the help of the original code (see Table 2).

Table 2. Summary of the reproduction efforts

	Fully	**Partial**	**No**
Raw data provided	x		
Cleaning code provided	x		
Analysis data provided	x		
Analysis code provided	x		
Reproducible from raw data	x		
Reproducible from analysis data	x		

We found two minor coding errors in the replication package while reproducing the study. First, we noticed that in the code entitled “Exp3.analyze.Rmd,” there is a variable name "idea_id" instead of “idea_text_id.” In the code entitled “Exp5.analyze.Rmd”, the code asks for the file “Exp5.rating.link_High.csv”, but the correct name of the file is “Exp5.rating.link_EmpathyHigh.csv”. We provide the correct code in Appendix 1.

2.1 Discrepancies Between Pre-analysis Plan and Article

Out of the six experiments, two (Experiments 1 and 3) were not pre-registered, and four (2A, 2B, 4, and 5) were pre-registered without a detailed Pre-Analysis Plan. We found links to AsPredicted.org pre-registrations of all four experiments within the paper. The registrations contain only basic information about the studies without details about the exclusion criteria (how nonsensical /out-of-context/ responses will be identified), variables (how creativity scores will be computed), analyses (significance level is not pre-registered), and hypotheses testing (which statistics will be considered when deciding about the support for the hypotheses). The authors just described the experimental

conditions (e.g., ChatGPT-assisted), dependent variable (creativity), provided the name of the main analysis (e.g., ANOVA with contrasts), information about the target samples size, and general information about which data will be excluded (e.g., nonsensical responses; who fail the attention check). We verified that authors consistently used the analyses specified in the pre-registrations to test their hypotheses. They also used other non-pre-registered tests for their supplementary analyses, manipulation checks, and robustness checks. Nevertheless, the pre-registered and non-pre-registered analyses are not distinguished within the manuscript. We found that the authors applied non-preregistered exclusion criteria when cleaning the data, as they excluded respondents who did not complete the survey (even if they provided their ideas) and who did not provide information about their gender or age. Although including only fully completed survey responses can be reasoned, applying this non-preregistered criterion is a minor deviation from the protocol. The real sample sizes slightly differ from the pre-registered sample sizes, which is common for online experiments with panels of respondents, as many people complete the survey in a very short time frame.

3. Robustness Reproduction

The authors already conducted several robustness checks and replications to ensure the reliability and validity of their findings. They replicated the effect of ChatGPT across six experiments using different creative tasks and contexts, comparing ChatGPT-assisted performance not only to web-search-assisted conditions but also to human-only conditions, and validated creativity ratings with both lay judges and expert judges, yielding consistent results. Mediation analysis (part of Experiment 3) helped explain the mechanisms behind the tested effect, showing that ChatGPT's impact on creativity was attributed to improved idea exposition and articulateness. Furthermore, the authors examined potential moderators, such as high versus low task constraints (Experiment 4) and empathy requirements (Experiment 5), finding that the ChatGPT-assistance effect persisted in all cases. Finally, alternative explanations, such as perceived conversational experience, task engagement, or affect, were ruled out through supplementary analyses. In Experiments 2-5, the authors also asked participants about how much they modified the ChatGPT outputs when providing answers to the creativity task. In a supplementary

analysis, the authors demonstrated that there was no significant difference in the assessment of creativity of responses among participants who self-reported that they did not modify their answers at all and those who did (see Supplementary Note H of the original manuscript). Together with the results of experiment 2B, this suggests that ideas generated by ChatGPT are more creative than those generated by humans, regardless of whether humans contribute to ChatGPT's outputs in any way. Nevertheless, we believe that self-reporting on the modification of answers provided by ChatGPT is not the most accurate indicator of the extent to which respondents actually modified ChatGPT-generated answers. Paid respondents from the panel may have been reluctant to admit that they did not follow the instructions exactly and simplified their work by simply copying and pasting the ChatGPT answer. Moreover, the binary variable (did not modify at all vs. modified) also does not capture the varying efforts that respondents put into modifying ChatGPT's output. Data from the experiments provides information about response time, which may be a more accurate indicator of the degree of ChatGPT involvement or personal effort. It can be assumed that respondents with longer response times devoted more human effort to working with ChatGPT than those with shorter response times. Therefore, as a robustness check, we decided to take response time into account and test whether the effort invested in modifying ChatGPT outputs, measured as the time spent on the experiment, has an incremental benefit for creativity. We examined whether the time invested in the experiment is related to creativity ratings in the ChatGPT-assisted condition and whether it moderates (i.e., strengthens) the effect of ChatGPT use on creativity. For our robustness check, we used data from Experiment 3. To avoid bias caused by extreme response times, we recoded response times using a logarithmic (ln) transformation. To interpret the moderation effect more clearly, we also performed a Z-transformation on the logarithm of the response time. We provide the full code and results in supplementary online materials at OSF.io.

First, we compared the response times of respondents in the ChatGPT-assisted (M = .042, SD = 0.945) and Web-search-assisted (M = -.034, SD = 1.050) conditions using a Welsch t-test, $t(191.15) = 0.567$, $p = .571$, $d = .08$, 95%Ci_d [-.20; .36]. The results showed no significant difference between the groups, indicating that the respondents who were expected to modify the ChatGPT outputs mostly did not only copy and paste the ChatGPT

outputs, but also invested similar effort in the task as respondents who were expected to use a web search.

We then analyzed the relationship between the response time and creativity within the subset of respondents assigned to the ChatGPT condition. According to our analysis, the response time did not correlate with the creativity ratings, r = -.079, p = .449, indicating that the time spent modifying the ChatGPT output does not lead to more creative solutions. This supports the conclusion that human effort has no incremental benefit on the creativity of ideas generated by AI.

To test the moderating effect of response time on the effect of experimental condition on creativity, we used two consecutive ANCOVAs with experimental condition (ChatGPT vs. Web-search-assisted) as a factor, response time as a covariate, and creativity as a dependent variable. The first ANCOVA, $F(2, 191) = 3.387$, p = .036, showed a significant effect of ChatGPT condition, b = 0.243, S.E. = 0.094, p = .011, when controlling for the response time. Response time did not significantly influence creativity, b = 0.015, S.E. = 0.047, p = .757. The second ANCOVA, with an interaction between experimental condition and response time, showed similar results, $F(3, 190) = 2.913$, p = .036. ChatGPT condition predicted significantly creativity, b = 0.243, S.E. = 0.094, p = .011, while response time, b = 0.072, S.E. = 0.063, p = 0.251, and interaction, b = -0.132, S.E. = 0.095, p = .166, had no significant effect on creativity. This result provides further evidence that the effect found by Lee and Chung (2024) describes the differences in creativity between humans and AI, and that human input makes no measurable incremental contribution to creativity over the large linguistic model. Our results report the average effect across the entire sample and do not exclude the possibility that, in the case of some individuals, human effort made a positive or negative contribution beyond the results generated by the GPT chatbot.

4. Conclusion

We successfully reproduced the key analyses presented in the Lee and Chung (2024) paper, supporting the main claim that using ChatGPT assistance significantly increases the creativity of responses to problem-solving tasks compared to web-search assistance or no assistance. We computationally reproduced all reported results using the original dataset and code, and successfully recreated the analyses comparing creativity in ChatGPT-assisted and Web-search-assisted conditions in Experiments 1 and 3 with newly written code. Therefore, we found that the study is fully reproducible from the raw data, although the non-pre-registered and non-described cleaning steps, as well as the missing codebooks, make reproduction difficult.

Our robustness check using response time as a proxy for human effort revealed no incremental benefit of human modification on creativity beyond ChatGPT's contribution. This finding strengthens the interpretation that the observed effect reflects differences between human and AI-generated ideas rather than additive human input.

Future replication efforts could extend our work in two directions. First, robustness checks could be applied to other experiments in the original study to verify whether the absence of incremental human contribution generalizes across various tasks and contexts. Second, direct replications with new data could explore conditions under which humans might enhance perceived creativity beyond ChatGPT outputs. Such studies could examine interpersonal differences (e.g., cognitive styles, expertise), motivational factors, further task characteristics, or measures of creativity. These extensions would deepen our understanding of human–AI collaboration and its potential limits.

References

Brodeur et al. (2024). Reproduction and replication at scale. Nat Hum Behav 8, 2–3 (2024). https://doi.org/10.1038/s41562-023-01807-2

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), 1906-1914.

Appendix 1: Repairing minor coding errors in Experiment 3 and 5

1.) Code from “Exp3.analyze.Rmd”

```
participant_to_idea_id = read.csv("./Exp3.rating.link.csv") %>%
  select(-idea_text_id) %>%
  rename(idea_id =X)
```

2.) Code from “Exp5.analyze.Rmd”

```
participant_to_idea_id.EmpathyHigh = read.csv("./Exp5.rating.link_EmpathyHigh.csv")
%>%
 select(-idea_text_id) %>%
  rename(idea_id =X) %>%
  mutate(idea_id = paste0("H",idea_id))
```