INTERNATIONAL DESIGN CONFERENCE – DESIGN 2022
https://doi.org/10.1017/pds.2022.88

Inspirational Stimuli Improve Idea Fluency during Ideation: A Replication and Extension Study with Eye-Tracking

H. Dybvik [1,✉], F. G. Abelson [1], P. Aalto [1], K. Goucher-Lambert [2] and M. Steinert [1]

[1] Norwegian University of Science and Technology, Norway,
[2] University of California, Berkeley, United States of America

✉ henrikke.dybvik@ntnu.no

Abstract

We replicate a design ideation experiment (Goucher-Lambert et al., 2019) with and without inspirational stimuli and extend data collection sources to eye-tracking and a think aloud protocol to provide new insights into generated ideas. Preliminary results corroborate original findings: inspirational stimuli have an effect on idea output and questionnaire ratings. Near and far inspirational stimuli increased participants' idea fluency over time and were rated more useful than control. We further enable experiment reproducibility and provide publicly available data.

Keywords: design cognition, conceptual design, eye tracking, experimentation, replication

1. Introduction

An important part of science and experiments is reliability, repeatability (or replicability), and reproducibility—however, experiments are not replicated as often as they ought to be or fail to replicate for numerous reasons. Open Science Collaboration's recent effort to conduct 100 replications of systematically sampled psychology results in top-tier journals, produced significant results in 36% of the replication studies, which, compared to 97% significant results in original studies (Open Science Collaboration, 2015), we find shocking. Moreover, 32% of original results were not significant when combined with new data (Open Science Collaboration, 2015; Shrout and Rodgers, 2018). The reproducibility crisis, across scientific fields, is exacerbated by a publication bias towards statistically significant results and reluctance to publish replication studies (Field, 2018; Martin and Clarke, 2017; Shrout and Rodgers, 2018). Even studies of exemplary quality may have irreproducible results due to random or systematic error, replication is therefore not only an opportunity to improve reproducibility—it is necessary (Open Science Collaboration, 2015).

We want to minimize potential replication issues in design research and advocate for providing replication efforts with a positive connotation. To this end, we have replicated the focus and activity (i.e., experimental design and stimuli) of a design ideation study that used neuroscience methods and means of data collection (Goucher-Lambert et al., 2019), and extended the study by changing and adding new sources of data collection. The design ideation study, referred to as "the original study" throughout this article, conducted an experiment were participants lay supine in an fMRI and were tasked to generate ideas for 12 design problems assisted by word stimuli that were inspirational, either near- or far from the solution space, or that served as a control (Goucher-Lambert et al., 2019). The original study explored the impact of inspirational stimuli on design ideation, behavioral-, and neurological processes, and demonstrated that inspirational stimuli both near and far from the problem space increase idea fluency compared to control stimuli. Inspirational stimuli was most beneficial after

some time, it enabled participants with a higher idea output over time. Inspirational stimuli had a significant effect on subjective ratings of relevancy and usefulness of the stimuli, but not on quality and novelty of the ideas. fMRI data suggested two search strategies: In the positive strategy—the inspired internal search—participants recognize the inspirational stimuli as applicable to the design problem, and it activates brain regions associated with memory retrieval and semantic processing. The negative strategy—unsuccessful external search—participants continue searching the problem space for an inkling, and it increases activation in brain regions associated with directing attention outwards and visual processing. Control stimuli is consistent with the negative strategy, while near stimuli triggers the positive strategy. Far stimuli, depending on the actual distance from the problem space, exhibits features from both strategies.

The replication and extension study presented in this article originated from us possessing eye-tracking technology while simultaneously contemplating the nature of the original study's stimuli; specifically, when observing the different words within each stimuli we wondered whether any of them were more or less "inspirational" and whether participants paid more attention to them. Eye-tracking may provide insights into visual allocation across various stimuli, revealing potentially subconscious behavior during ideation. The present study was solely driven by a desire to investigate these questions with eye-tracking technology; it assumes the fMRI results' validity since it does not have access to an fMRI for verification purposes. Further, since original participants ideated silently there is no record of which ideas were produced, thus, it is unknown whether the resulting ideas were different; we wanted to address this as well, by adding a think aloud protocol. The eye-tracking and think aloud combination is particularly interesting to us since it could reveal exactly which words were used when producing specific ideas. An fMRI environment is restrictive, does not allow for a think aloud protocol without affecting data quality, and share few commonalities with practitioners' ideation. By moving the experimental design and task from the fMRI context to a conventional office-with-desk context we obtain a more realistic experimental context, and it becomes interesting to investigate if the number of generated ideas and participants subjective ratings hold true across contexts. We replicate the original experimental design and stimuli, change context from an fMRI to a conventional desk, and extend by adding eye-tracking and think aloud protocol as means of data collection.

This article describes the experiment briefly, including replicated- and new content, and adaptations required to include the new data collection sources properly. We further present preliminary results from analyzing the number of generated ideas and participants' subjective ratings of ideas, which largely corroborate original results. Eye-tracking data and recording of the think aloud protocol is currently under analysis and will be presented in future publications.

2. Background

2.1. Similar research

Eye-tracking as a research tool has gained popularity across several research fields the past 20 years (Carter and Luke, 2020). In design research eye-tracking is listed among the tools for studying design physiology and that it “gives insight into visual reasoning during a design task” (Gero and Milovanovic, 2020). A substantial amount of existing research uses eye-tracking as a tool to explore engineering and product design using image stimuli. A search for similar eye-tracking research was performed by querying "design ideation" AND "eye-tracking" in Google Scholar, and querying "eye-tracking" amongst the 48 citations of Goucher-Lambert et al. (2019). An extensive review of existing design research using neurophysiological and biometric measures, thus including eye-tracking, was also used to search for similar work (Borgianni and Maccioni, 2020). This search did not find any study using eye-tracking to explore the effects of inspirational word stimuli in design ideation, but examples of eye-tracking for other or similar design ideation tasks were found. Cao et al. (2018) also uses stimuli of varying distance from the problem space to explore difference between beginning and advanced design students during idea generation, but uses images as stimuli. Kwon et al. (2019) looked into the relation of eye movements and idea output (creativity) to an “alternative uses test (AUT)”, where participants are presented with 12 object images and get 2 minutes per object to name alternative uses of the object.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

This study shares similarities with the idea generation study by Colombo et al. (2020) in which AUT and eye-tracking explore the differences between designers and engineers.

2.2. Eye-tracking technology

Eye-tracking is a measure of eye movements, and thus gaze location over time (Carter and Luke, 2020). Recording of eye movements dates back to 1823, and the technology have seen vast improvements in recent years using video-based eye-trackers making it more affordable and accessible (Carter and Luke, 2020; Wade et al., 2005). There are two main types of video-based eye-trackers: table and head-mounted configurations (Carter and Luke, 2020). Head-mounted eye-trackers work by shining infrared light at the eye, and illuminating it without being visible to humans, resulting in a corneal reflection and bright pupil effect (Duchowski, 2017). The corneal reflection appears as a glint on the eye, and the bright pupil effect are both caused by the reflection of the infrared light and are recorded using eye-facing cameras. By using the location of the corneal reflection and the pupil center, software can calculate the gaze position after device calibration. Eye-tracking data are time series sampled at a given frequency yielding the gaze position (Carter and Luke, 2020). When the eyes fixate on a target over a period of time the gaze points can be aggregated into a fixation. Fixation length vary and are usually within the range of 180-330 milliseconds (Rayner, 2009). The rapid eye movements between fixations, happening while scanning the visual space and moving the eyes, are called saccades and during these the visual input is suppressed (Carter and Luke, 2020; Rayner, 2009).

3. Method

3.1. Experimental design and setup

3.1.1. Ideation task within a repeated measures experimental design

The task, and thus problems and words used in this experiment are the exact same as in Goucher-Lambert et al. (2019). Participants were tasked to develop as many ideas as possible for 12 different design problems and instructed to "thinking aloud" by briefly explaining their idea in a think aloud protocol. Each new idea was indicated by pressing the space bar. Five words were presented along with each problem. Reused words from the problem statement was presented in the Control condition. Words near or far from the problem space was used as inspirational stimuli in the Near and Far condition respectively. The 2 minutes ideation time per problem was divided into two blocks of 1 minute. The first block called Wordset1 displayed the three first words. The second block called Wordset2 displayed all five words. The 1-back memory task was performed between blocks. Participants completed a questionnaire—rating the usefulness and relevancy of the words presented, and the novelty (uniqueness) and quality of the solutions developed on a scale from 1 to 5—after each problem. The experiment follows a repeated measures design assigning participants to one of three counterbalanced groups of specific problem-condition pairs. The full experimental procedure with routines' timing is visualized in Figure 1. The fMRI-specific fixation cross routine, indicated by “+”, was kept to retain the original study's temporality, and had a random duration between 0.5 and 4 seconds.

3.1.2. Differences from original study

The main difference between the original study and this study was the use of eye-tracking technology. Originally, participants lay supine in an fMRI viewing stimuli on a monitor through a look out mirror attached to the head mounted coil. By using a response glove strapped to their right hand, participants could indicate new ideas with their index finger and provide questionnaire ratings with all five fingers. In this experiment participants sat in a chair in front of a monitor, equipped with a conventional computer mouse and keyboard to indicate new ideas and submit questionnaire ratings.

Participants were additionally tasked to think aloud which may impact the number of ideas produced as it may require more time to articulate an idea compared to only thinking of it.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

3.2. Hardware

The experiment was run on a conventional desktop computer with a 24 inch monitor along with the head-mounted eye-tracker from Pupil Labs (Kassner et al., 2014) with binocular setup (cameras on both eyes). Participants were placed in a chair approximately 70 cm from the monitor, see Figure 2. We weren't interested in sub-word accuracy, but rather areas, words, and patterns as a whole meaning that a higher accuracy obtained by a chin rest wasn't necessary. We believed a chin rest would feel restricting for participants during ideation, perhaps also increasing a Hawthorne effect or other expectancy biases, and thus chose to not use one. A USB-connected microphone was placed on a tripod in front of the participant to obtain high quality audio recordings. A conventional keyboard and mouse were used. Additional hardware specifications are listed below.

Hardware specifications:

- Desktop computer: Dell OptiPlex 7050, OS: Windows 10 Education 64-bit, CPU: Intel Core i7-7700 @ 3.60GHz, RAM: 32 GB
- Monitor: Dell UltraSharp U2412M, Size: 24” (61 cm), Resolution: 1920x1200 pixels, Refresh rate: 60 Hz
- Microphone: Zoom H1 Handy Recorder, f_s: 48 kHz, Bit rate: 16 bit, Channels: 1 (mono recording)
- Eye-tracker: Pupil Core, World cam. Resolution: 1280x720 pixels, fs: 30 Hz, Field of view: 99 degrees x 53 degrees, Eye cam. Resolution: 192x192 pixels, f_s: 120 Hz. Gaze Accuracy: 0.6 degrees, gaze precision: 0.02 degrees.

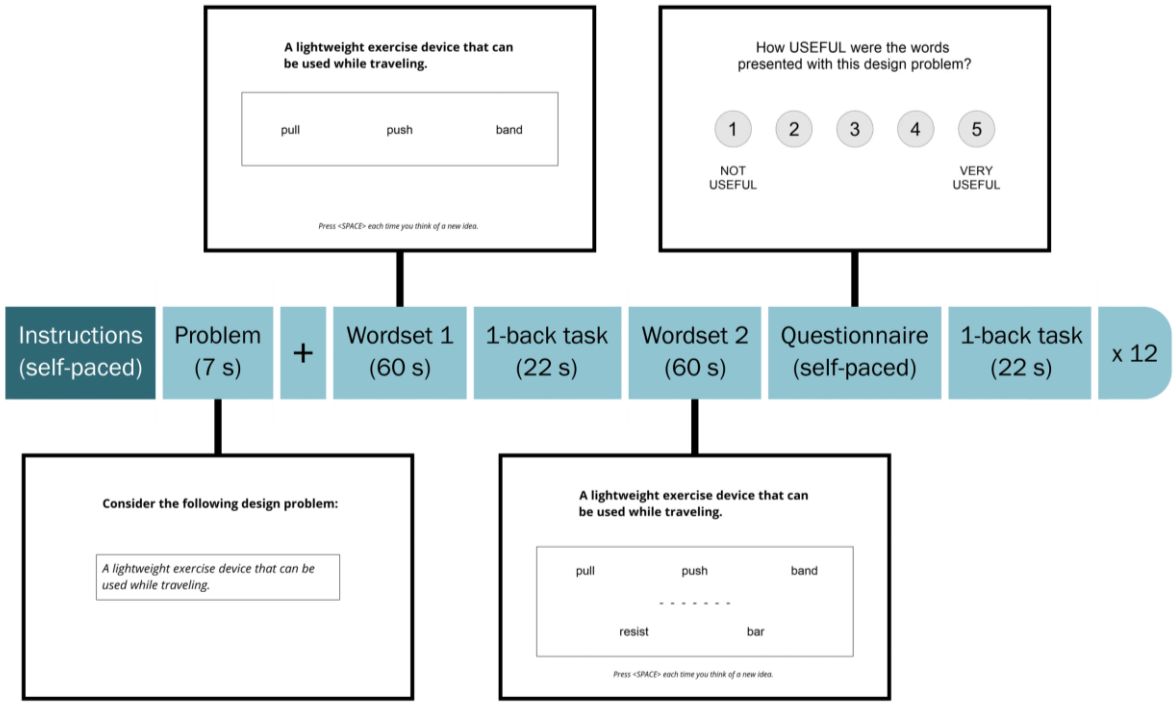

Figure 1. Experiment procedure. "+" indicates fixation cross. Instructions were shown once.

3.3. Software implementation

The experiment was programmed in the open-source software PsychoPy v2021.1.4 (Peirce et al., 2019) in contrary proprietary software E-Prime used originally. PsychoPy offers a graphical user interface and allows for running custom Python code. Most input data and visual design was retrieved from the published article. Some additional figures and information were obtained from original authors. All word stimuli were presented as black text on a white background in OpenSans font. Letter height were set to 5 percent of the screen height in PsychoPy which translates to 60 pixels on the monitor.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

3.3.1. Eye-tracking data collection, annotations and time synchronization

Eye-tracking data was collected with Pupil Capture. Pupil Network API was used to synchronize eye-tracking data, audio data, questionnaire responses, timestamped ideas, and stimuli annotations. The Network API control time synchronization of PsychoPy and Pupil Capture by sending a message over the API setting Pupil's clock to the global experiment clock in PsychoPy. Automatic data recording was implemented using the API, ensuring that Pupil Capture began recording once the PsychoPy experiment was launched.

3.3.2. Audio recording

Automatic sound recording of participants thinking aloud was implemented by using high-level functionality from module python-sounddevice to record, and saving functionality in WAV format from SciPy. Each recording was automatically saved with a filename with participant ID, problem ID and stimuli conditions.

3.3.3. Surface tracking

We used Pupil's Surface Tracker plugin to record the gaze of the participants relative to the monitor, not only the video frame. By fixing AprilTags (small binary markers) on the bezel of the monitor the plugin can map out the planar monitor surface, thus marking the exact size of the monitor in recording software. We designed and 3D printed custom monitor mounts to ensure no changes in marker setup.

Figure 2. Left) Monitor with apriltags, Pupil Core to the right, and microphone in the middle. Right) Experimenter monitors eye-tracking real time during experimental run.

3.4. Participants

24 healthy adults (18 male/6 female, ages 23-35, mean = 25.8 yrs., SD = 2.9 yrs.,) participated in the study. 22 were right-handed and 2 left-handed. None of the participants were native English speakers, and none wore glasses to not interfere with the head-mounted eye-tracker. 8 participants used lenses. Participants were recruited through internal channels and contacts at relevant departments at the Norwegian University of Science and Technology (NTNU)—the Department of Mechanical and Industrial Engineering (MTP) and the Department of Design (ID)—to ensure a similar educational background as original participants. All participants were graduate level students or higher (minimum 4th year MSc, PhDs), with more than half of the participants being final year Master students. No monetary compensation was given.

3.5. Experiment procedure and calibration

After providing informed consent participants received general information in Norwegian about the experiment, its procedure, and the task. The eye-tracker was then correctly positioned on the participant before calibration of the eye-tracker. A 3D calibration was performed following manufacturers' "Best Practices" (Pupil Labs, 2021). Afterwards, the experiment started by showing additional information, before proceeding to explaining the design ideation task again and the 1-back task. Participants were

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

sequentially assigned to groups in order A, B, C. After completing the approximately 1 hour long experiment participants answered a demographic survey.

3.6. Knowledge from pilot participants

The experiment was piloted, following several procedures by van Teijlingen and Hundley (2001) to remove unexpected technical bugs or ensure clear task description. The experiment was conducted as if would be for the actual participants, the session was timed, feedback from participants to identify potential ambiguities was obtained afterwards, and eye-tracking data quality was inspected. This induced two experimental changes: 1) The initial chairs height led participants to angle their head which degraded eye-tracking data quality, and thus the chair was changed to one with an appropriate height relative to the table. 2) Priming the participants. Participants were unsure of what "rules" that applied during the design ideation. For example, did it only have to be new ideas or realistic ideas? The study's primary purpose wasn't a quality evaluation of generated ideas. The manuscript briefing participants initially was thus written informing participants about ideating freely and without any constraints, and a change in the written instruction in PsychoPy was made. Changes were iteratively implemented before commencing with actual participants.

3.7. Data Analysis

To summarize, the following data modalities were collected during the experiment: eye-tracking data, audio recordings, number and timing of generated ideas, and a questionnaire. The scope of this article is to analyze the number of ideas and subjective ratings from the questionnaire. Analysis of eye-tracking data and audio data is not within the scope of this article and will be published later.

Compared to the original study we hypothesize that this study will result in a similar number of ideas in order of magnitude, perhaps fewer du to having to think aloud and cultural factors. The subjective ratings and differences between ratings will be similar.

Questionnaire: Differences in subjective ratings between conditions were assessed with Friedman's test, a non-parametric test, suitable for ordinal data such as ratings on a 1-5 scale. Post hoc pairwise comparisons were assessed with Wilcoxon signed rank test with a Bonferroni correction for multiple comparisons (Field, 2018). Hedges' g was used as effect size.

Idea generation: For idea generation analysis the number of ideas were aggregated per stimuli and wordset. Differences in the number of ideas generated between conditions was assessed with one-way repeated measures ANOVA, suitable for continuous variables (Field, 2018). ANOVAs assumptions of sphericity and normal distribution of the data were assessed. The Control-Wordset1 contrast failed the normality test, but since ANOVA is relatively robust against normality violations, and all other data exhibited both sphericity and normality, we continued with the analysis. Partial-eta squared (η^2) was used as effect size. Statistics were performed in Python with Pingouin (Vallat, 2018), Pandas (McKinney, 2010), Seaborn (Waskom, 2021), and Matplotlib (Hunter, 2007). The significance level was set at $p<0.05$ for all tests.

4. Results and discussion

4.1. Questionnaire results

There was a highly significant difference between conditions for relevancy and usefulness, insignificant difference for novelty, and a $p = 0.051$ for quality. See Table 1 and Figure 3.

Post hoc pairwise comparisons are listed in Table 2. There was a significant difference in relevancy between Control and Near, and between Far and Near with Near being more relevant than both Control and Far. There was a significant difference in usefulness between Control and Far, and between Control and Near, with the inspirational stimuli conditions being more useful than control. Moreover, the difference between Far and Near reached a significance level of $p=0.05$, which we find interesting and interpret as a strengthening indication of that Near was more useful than far. There was not a significant difference between conditions for Novelty, indicating that participants did not consider their ideas to be more novel in either condition. There was one significant difference for Quality between Far and Near,

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

i.e., participants thought they produced ideas of higher quality in Near. These results exhibit similar trends to the original study with one exception: relevancy. Participants in this study rated Control to be less relevant than both Near and Far inspirational stimuli, whereas the original participants thought Control was more relevant than Near and Far. The experimenter noted during the experiment and in post experimental feedback that participants were unsure of what relevancy rating to give control stimuli, which may explain the difference. This ambiguity could potentially have been removed by clarifying whether to rate the relevancy of the words based on their relevancy for solving the problem or being related to the problem. Participants being non-native English speakers may also affect their interpretation of the question and the word "relevance", as it was read in English but processed and evaluated in Norwegian. Near and far inspirational stimuli was more useful than control, which corroborates the results from the original study. Near was close to significantly more useful than far in this study with a multiple-comparisons-corrected of 0.05, for which the original study reported a $p<0.01$. We don't know if this value is corrected for multiple comparisons or not, but if not, this might explain the discrepancy since this study's uncorrected p-value was 0.017. The novelty ratings corroborate the insignificant differences of the original study. Even though the trends were similar for questionnaire ratings, overall mean values for novelty and quality were lower. This can indicate lower confidence in the solutions generated by this study's participants.

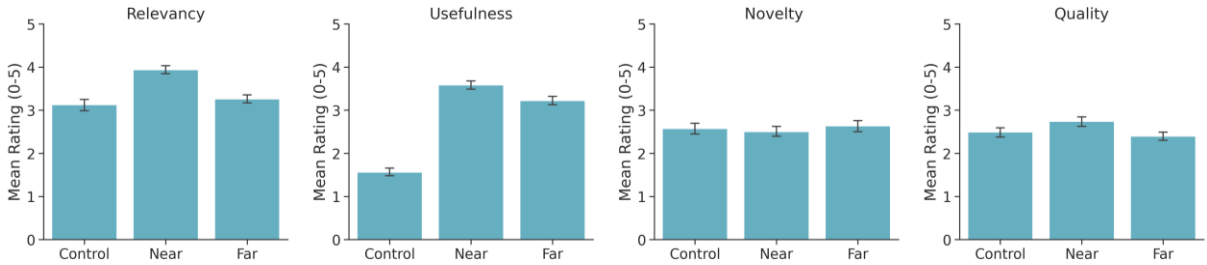

Figure 3. Mean ± 1 SE for participants subjective ratings

Table 1. Results subjective variables

	Control		Near		Far				
Variable	***M***	***SD***	***M***	***SD***	***M***	***SD***	***DOF***	***χ2***	***p***
Relevancy	3.12	1.35	3.94	0.92	3.26	0.94	2	16.587	<0.001**
Usefulness	1.56	0.83	3.58	0.94	3.22	0.99	2	39.758	<0.001**
Novelty	2.57	1.17	2.5	1.14	2.64	1.27	2	2.987	0.225
Quality	2.49	1.14	2.74	1.11	2.4	0.92	2	5.945	0.051

*p < 0.05, **p < 0.01.

Table 2. Pairwise comparisons subjective variables

Variable	**Between**		***W***	***p***	***Corr. p***	***Hedges' g***
Relevancy	Control	Far	126	0.726	1.000	-0.148
	Control	Near	51	0.015	0.044*	-0.886
	Far	Near	3	<0.001	<0.001**	-1.034
Usefulness	Control	Far	0	<0.001	<0.001**	-2.694
	Control	Near	0	<0.001	<0.001 **	-3.588
	Far	Near	35.5	0.017	0.05	-0.606
Novelty	Control	Far	73	0.886	1.000	-0.072
	Control	Near	89	0.362	1.000	0.088
	Far	Near	37.5	0.208	0.625	0.160
Quality	Control	Far	61	0.734	1.000	0.121
	Control	Near	64.5	0.133	0.398	-0.301
	Far	Near	20	0.008	0.023*	-0.428

*p < 0.05, **p < 0.01. Uncorrected p-value included for understanding of the effect of multiple comparisons correction.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

4.2. Idea generation results

The number of ideas produced in Wordset1 was significantly higher than in Wordset2 for all conditions (Control: $t(23) = 7.250$, $p < 0.001$, Near $t(23) = 5.152$, $p < 0.001$, Far: $t(23) = 7.052$, $p < 0.001$). See Figure 4 which also illustrates the differences between participants in the numbers of ideas produced. The number of ideas generated in each condition plotted over time in Figure 5 exhibits a similar shape as in the original study, but with an approximate 10 second temporal delay. This may have been caused by participants both being native English speakers, the think aloud protocol, or a combination thereof with or without other influencing factors.

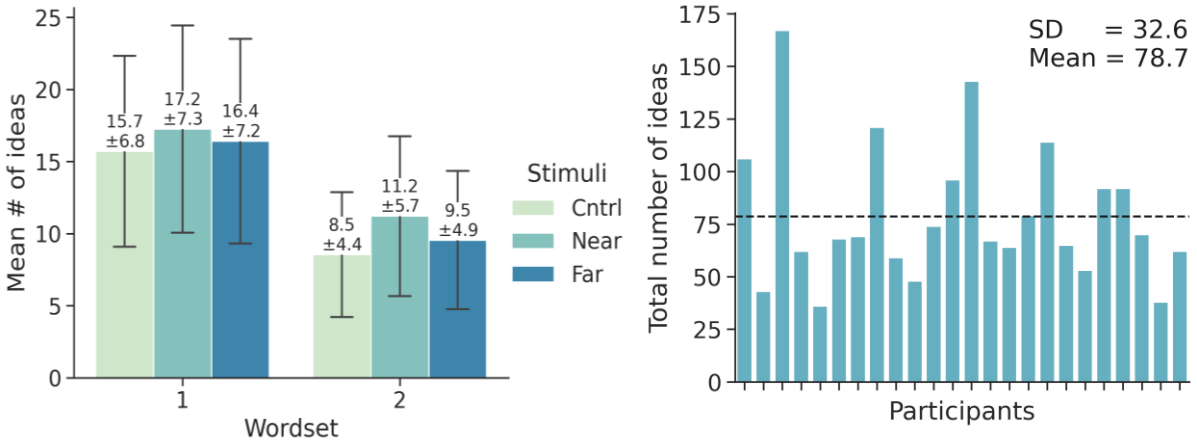

Figure 4. The number of ideas produced across stimuli and participants. Mean ± SD are on aggregated ideas across all wordset-stimuli combinations.

In Figure 4 and Figure 5 Near stimuli appears to generate more ideas than Far stimuli, again generating more ideas than Control stimuli—for both wordsets. However, there was not a statistically significant different number of ideas produced between conditions for Wordset1 ($F(2, 46) = 2.241$, $p = 0.118$, $\eta 2 = 0.089$). This corroborates the original study's results. The number of ideas was significantly different between conditions for Wordset2. ($F(2, 46) = 10.316$, $p < 0.001$, $\eta 2 = 0.310$). Post hoc pairwise comparisons for Wordset2 reveals a significant difference between Control and Near, and between Far and Near. See Table 3. This indicates that inspirational stimuli help participants retaining a higher idea output throughout the ideation session. It also interesting to note that ideas produced with Far inspirational stimuli was not significantly different from Control, contrary to the original study finding this contrast significant. The original study also resulted in a significant Control-Near contrast, but their Far-Near contrast was not significant at $p<0.05$, although it was close with a $p=0.06$. Both studies' mean values of idea output were indeed Near > Far > Control. Overall, these results indicate that inspirational stimuli nearer to the problem space facilitates idea generation. Moreover, original results have largely been corroborated.

Table 3. Pairwise comparisons number of ideas

Wordset	Between		*T*	*DOF*	*p*	*Corr. p*	*Hedges' g*
1	Control	Far	-1.026	23	0.315	0.946	-0.099
	Control	Near	-2.157	23	0.042	0.125	-0.215
	Far	Near	-1.069	23	0.296	0.888	-0.112
2	Control	Far	-1.906	23	0.069	0.208	-0.211
	Control	Near	-3.968	23	0.001	0.002**	-0.515
	Far	Near	-2.908	23	0.008	0.024*	-0.309

*p < 0.05, **p < 0.01. Uncorrected p-value included for understanding of the effect of multiple comparisons correction.

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

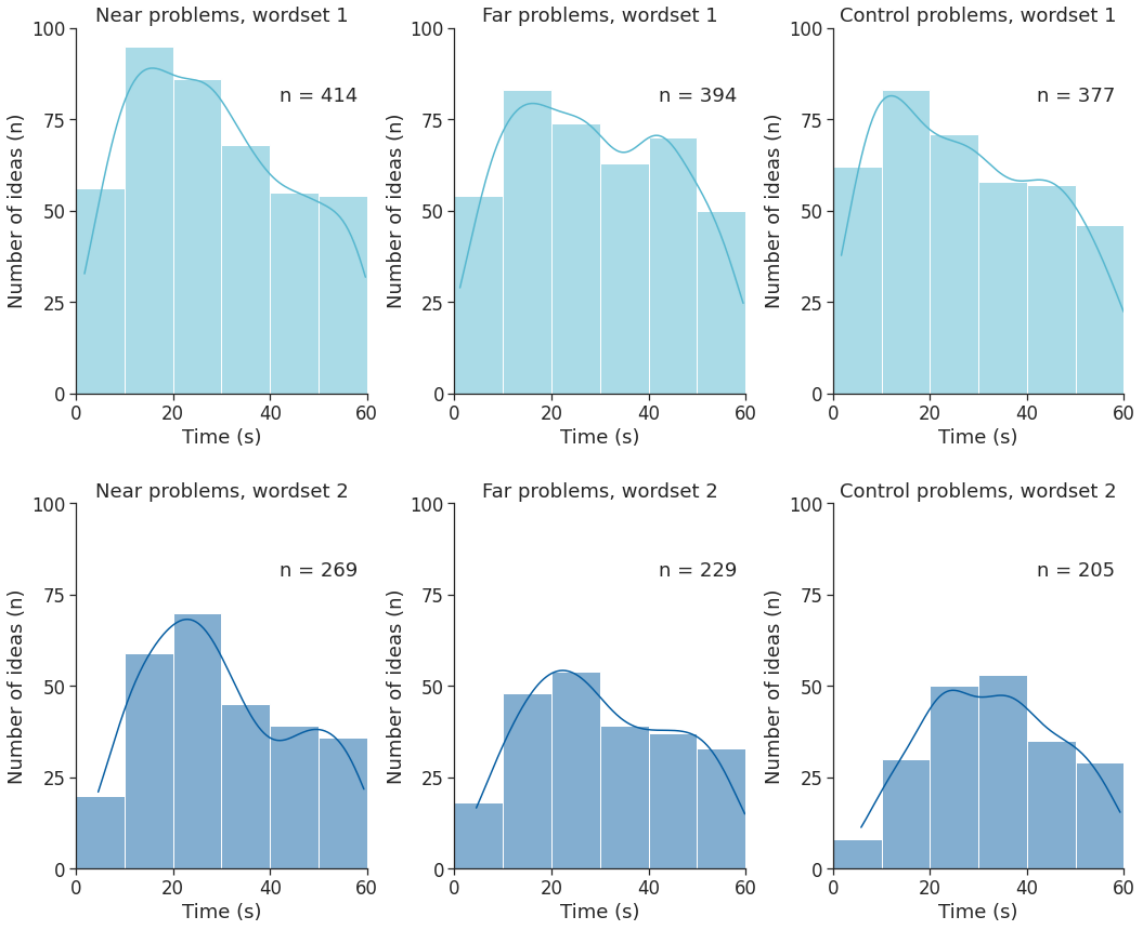

Figure 5. Histograms with number of ideas over time generated across all conditions, binned into bins of width 10 second. All histograms are overlaid with a kernel density estimate (KDE).

4.3. Further work

We are currently analyzing the eye-tracking data and audio recordings to get further insights to what kinds of ideas that were produced for the different problems across the different conditions. The audio recordings aren't published due to privacy considerations and a transcription will therefore be completed, and eventually publicly available. Further replication of the experiment may bring insight into potential differences in results due to culture and/or nationality, and is also of interest to further understand the effects of inspirational stimuli on idea generation.

5. Conclusion

This article described the replication and extension of a design ideation experiment with and without inspirational stimuli. Eye-tracking technology and a think aloud protocol was added to provide new insights into generated ideas. Preliminary results presented here corroborates the original study's results, inspirational stimuli influence idea output and questionnaire ratings. Participants produced more ideas over time when aided by inspirational stimuli, and rated inspirational stimuli as more relevant and useful than control. Future work will analyze eye-tracking data and audio recordings. The experiment and the data, and code are publicly available for reproducibility purposes.

Acknowledgement and Data availability

This work has been published in a master thesis.

Raw data, analysis code, results, and PsychoPy experiment code are all publicly available: code repository (Abelson, 2021), pre-processed data: (Abelson et al., 2021a), and raw eye-tracking data (Abelson et al., 2021b).

https://doi.org/10.1017/pds.2022.88 Published online by Cambridge University Press

References

Abelson, F.G., 2021. Code Repository for Design Ideation Experiment (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5130090

Abelson, F.G., Dybvik, H., Steinert, M., 2021a. Dataset for Design Ideation Study. DataverseNO. https://doi.org/10.18710/PZQC4A

Abelson, F.G., Dybvik, H., Steinert, M., 2021b. Raw Data for Design Ideation Study. https://doi.org/10.21400/7KQ02WJL

Borgianni, Y., Maccioni, L., 2020. Review of the use of neurophysiological and biometric measures in experimental design research. AI EDAM 34, 248–285. https://doi.org/10.1017/S0890060420000062

Cao, J., Xiong, Y., Li, Y., Liu, L., Wang, M., 2018. Differences between beginning and advanced design students in analogical reasoning during idea generation: evidence from eye movements. Cogn Tech Work 20, 505–520. https://doi.org/10.1007/s10111-018-0477-z

Carter, B.T., Luke, S.G., 2020. Best practices in eye-tracking research. International Journal of Psychophysiology 155, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Colombo, S., Mazza, A., Montagna, F., Ricci, R., Monte, O.D., Cantamessa, M., 2020. NEUROPHYSIOLOGICAL EVIDENCE IN IDEA GENERATION: DIFFERENCES BETWEEN DESIGNERS AND ENGINEERS. Proceedings of the Design Society: DESIGN Conference 1, 1415–1424. https://doi.org/10.1017/dsd.2020.161

Duchowski, A.T., 2017. Eye-tracking Methodology, 3rd ed. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-57883-5

Field, A., 2018. Discovering statistics using IBM SPSS statistics, 5th edition. ed. SAGE Publications, Thousand Oaks, CA.

Gero, J.S., Milovanovic, J., 2020. A framework for studying design thinking through measuring designers' minds, bodies and brains. Design Science 6. https://doi.org/10.1017/dsj.2020.15

Goucher-Lambert, K., Moss, J., Cagan, J., 2019. A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli. Design Studies 60, 1–38. https://doi.org/10.1016/j.destud.2018.07.001

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Kassner, M., Patera, W., Bulling, A., 2014. Pupil: an open source platform for pervasive eye-tracking and mobile gaze-based interaction, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct. Association for Computing Machinery, New York, NY, USA, pp. 1151–1160. https://doi.org/10.1145/2638728.2641695

Kwon, E., Ryan, J.D., Bazylak, A., Shu, L.H., 2019. Does Visual Fixation Affect Idea Fixation? Journal of Mechanical Design 142. https://doi.org/10.1115/1.4045600

Martin, G.N., Clarke, R.M., 2017. Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. Frontiers in Psychology 8.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. Presented at the Python in Science Conference, Austin, Texas, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. https://doi.org/10.1126/science.aac4716

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. PsychoPy2: Experiments in behavior made easy. Behav Res 51, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pupil Labs, 2021. Best Practices - Tips for conducting eye-tracking experiments with the Pupil Core eye-tracking platform. [WWW Document]. Pupil Labs. URL https://docs.pupil-labs.com (accessed 4.27.21).

Rayner, K., 2009. Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology 62, 1457–1506. https://doi.org/10.1080/17470210902816461

Shrout, P.E., Rodgers, J.L., 2018. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. Annu. Rev. Psychol. 69, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Vallat, R., 2018. Pingouin: statistics in Python. Journal of Open Source Software 3, 1026. https://doi.org/10.21105/joss.01026

van Teijlingen, E.R., Hundley, V., 2001. The importance of pilot studies.

Wade, P. of V.P.N., Wade, N., Tatler, B.W., Tatler, L. in P.B., 2005. The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research. Oxford University Press.

Waskom, M., 2021. seaborn: statistical data visualization. JOSS 6, 3021. https://doi.org/10.21105/joss.03021