nature research

Corresponding author(s):   DBPR

Last updated by author(s): 2022-1-15

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For data collection in the lab studies, we used Qualtrics (accessed in 2016-2021) and Google Sheets (accessed in 2016 and 2021). The data from the field studies were collected from the employer. The employer used a proprietary software called InnoStreams to collect the ideas generated by teams (beta release, version 0.8). In the final two field studies, the employer used Qualtrics (accessed in 2019 and 2022) to collect and score ideas. The first batch of lab data collection used WebEx for virtual interaction (versions 36.6 -36.9). The second batch of data collection used Zoom (version 3.3). The virtual lab study used Zoom (versions 5.0.0 and 5.0.1). Engineers in the field studies used Webex (Portugal: 38.3, Finland and Israel: 38.6, Hungary: 39.3, India: 40.1).
Data analysis	We used R (4.0.1) and Python (3.8.1) to analyze ideas. OpenFace (2.2.0) was used to identify eye gaze. LIWC2015 (v1.6) was used to investigate language usage.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data collected by the research team reported in the main text and the supplementary materials is available on Research Box, https://researchbox.org/282&PEER_REVIEW_passcode=GHJKW, except the videos and recordings of participants because we do not have IRB permission to share

participants' voices or faces. The cleaned summary data is available in the same Research Box for the field studies, but the raw data must be kept confidential, as these data are the intellectual property of the company. The Linguistic Analysis database is available at https://liwc.wpengine.com/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences   ☒ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In all studies, participants were randomly assigned into a group, and the group was randomly assigned to work together in-person or virtually. Participants first generated ideas and then selected their best idea. Across all studies, we counted the number of ideas generated, and counted the number of creative ideas generated (using scores from judges). For idea selection, we investigated the score of the selected idea (and compared it to the score of the top idea).
Research sample	Our research samples included undergraduates from two private universities in the United States (batch 1: 202 females, 95 males, Mage = 26.1; SDage = 8.61; three participants did not complete the survey and thus are missing demographic information, batch 2:177 females, 119 males, 2 nonbinary, Mage = 23.5, SDage = 7.09; we are missing demographic and survey data from four participants, virtual study: 76 male, 174 female, 4 non-binary, and 2 undisclosed, Mage = 24.7, SDage = 8.19) as well as adult engineers. We do not have any demographic information from the engineers, as data were collected anonymously. During follow-up surveys, we also collected data from MTurk. These samples are not nationally representative. However, we chose these populations because they are relevant to our context of remote work. Both college undergraduates and engineers commonly engage in idea generation and selection. Further, both populations are familiar with videoconferencing technology and thus were particularly suitable for this study on virtual collaboration. In addition, both populations have engaged or will engage in remote work. Lastly, our samples are not exclusively WEIRD (western, educated, industrialized, rich, and democratic) as we analyze data from Hungary and India.
Sampling strategy	We used convenience sampling. No sample size calculations were performed. Given that the effect of virtual interaction on idea generation had never been tested, we had nothing to base an effect size estimate on. Thus, in the first batch of data collection in the lab, we selected a sample size that balanced reaching sufficient power for a medium to small effect with data collection efficiency. We initially analyzed this data using a Poisson distribution, and determined that 75 per cell was sufficient sample size. We then used the same sample size for our second batch of data collection. After data collection, we discovered that our count data (number of ideas and creative ideas generated) in our studies are "overdispersed" and thus, strictly speaking, violate one of the Poisson model's assumptions. We identified an alternative model that includes an overdispersion parameter: the negative binomial regression. However, the negative binomial regression decreases power by estimating an additional parameter and thus results in greater Type II error (Sturman, 1999). Indeed, a power analysis confirmed that each batch of data collection was underpowered for a negative binomial model: for our effect size, we would need 116 per cell for 80% power. Since our lab studies are stimulus replicates, we addressed this issue by combining these two batches of data, yielding 150 per cell and 89% power. For the field study, we were at the mercy of our corporate partner and its willingness to dedicate resources to a field study. We expressed interest in collecting at least 1,000 participants and indicated that we would prefer to receive data from as many workshops as they would be willing to run (given the additional noise that a field study creates and the need to drop groups that had technology failure). They were able to provide us over 1,600 workshop participants. A power analysis determined that we had 88% power with our final sample size of 745 groups.
Data collection	The research assistants who collected the data in the lab were blind to hypotheses. They were not present in the room while participants worked together. Lab study 1: Participants were recorded using an audio recorder, generated and selected ideas on Google Sheets, and responded to surveys on Qualtrics. WebEx was used in the virtual condition. Lab Study 2: Participants generated and selected ideas and responded to surveys on Qualtrics. They indicated their memory of the room on a sheet of paper and then listed these items in a Qualtrics survey. Video recordings were taken from the laptop camera and OpenFace was used to identify the gaze directions. Zoom was used in the virtual condition. Skype was used to share the task screen of the writer in both conditions. Field Study: The employer conducted field studies using Innostreams and Qualtrics. The engineers interacted on WebEx. The research assistants were not blind to hypothesis and were in the rooms to answer questions about the workshop and aid with technology issues. Group Size Survey: The employer collected data using Qualtrics. Screen Size Study: Participants on the Amazon Mechanical Turk platform responded to a Qualtrics survey. Virtual Study: All Participants interacted on Zoom and were recorded using Zoom. Participants generated and selected ideas on Google Sheets, and responded to surveys on Qualtrics. the research assistants were blind to hypothesis.
Timing	Lab Study 1: 6/28/16 - 10/6/16; Lab Study 2: 5/4/18 - 10/3/18; Field Study Poland: 3/14/18 - 3/16/18; Field Study Portugal: 3/19/18 - 3/20/18; Field Study Finland: 6/18/18 - 6/20/18; Field Study Israel: 6/21/18; Field Study Budapest: 3/18/19 - 3/21/19; Field Study India: 1/19/20 - 1/24/20; Field Study Finland Scoring: 1/20/22-2/11/22; Group Virtual Lab Study w/ Columbia Students: 2/12/21 - 3/31-21; Group Virtual Lab Study w/ Stanford: 1/4/21 - 5/12/21. Employer Group Size Survey: 11/10/21-11/16/21. MTurk Screen Size Survey: 11/10/21-11/12/21.
Data exclusions	No participants were excluded in Lab Study 1. In Lab Study 2, we excluded groups that contained any of 18 participants we recruited from craigslist after students reported feeling uncomfortable. We a priori excluded any pairs who experienced technical difficulties (e.g., screen share issues, audio feedback, or dropped video calls). In the field studies, we excluded 32 pairs due to technical difficulties, 21 pairs due to noncompliance, and 18 teams because they arrived late or left early. We also excluded 5 teams who had participated before and 15 three-person teams (when the session size contained an odd number of people). For the Finland idea

scoring, engineers were instructed to submit the survey a maximum of 3 times. We only analyzed the first three submissions, and we excluded "bad actors" who submitted 10 or more surveys. In the MTurk Screen Size Survey, we excluded anyone who indicated that they were not using videoconferencing at work in the screener (162 participants out of 400). In the video analysis, we a priori excluded nine videos because they were not saved, six videos because they cut off participants' eyes, four videos because they were too dark to reliably code, and two videos because they were corrupted and could not load. After coding, we excluded videos where at least 5 judges said that they couldn't hear or see the face of the participant (6 muted videos, 7 videos with sound). For eye gaze, we did not analyze eye gaze of 27 participants (out of 302) for the following reasons: Nine videos were not saved, six videos cut off participants' eyes, four videos were too dark to reliably code, two videos were corrupted and could not load, two videos contained participants with glasses that resulted in eye gaze misclassification, two videos (one team) did not have their partner to their right, and two videos were misclassified by OpenFace. For the virtual-only study, during data collection, we learned that some of the participants recruited using the student pool (university in the northeast) included community members not officially affiliated with the university. We a priori decided to drop those participants (4 groups) from analysis and continued collecting data until we reached the predetermined sample size. We also excluded 4 groups who experienced technical issues (as preregistered).

Non-participation

For the lab study, no participants dropped out. For the field studies, because we collected no information about the people attending the session run by their employer, we have no information on dropout (i.e., people who attended the session but didn't participate in the workshop).

Randomization

For the lab studies, because an elaborate set up is needed for each condition, we alternated conditions (e.g., group 100 was face-to-face, group 101 was in-person...). For the field studies, the employer randomly assigned engineers into pairs and randomly assigned pairs into condition (in person vs. virtual) by passing out shuffled nametags with a number and an A or a B (e.g., 100A) to all engineers. Each number represented a pair (e.g., 100A and 100B were assigned to work together), and all A's were the typists. Odd-numbered pairs were assigned to work together virtually, and even-numbered pairs were assigned to work together in person.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

See above.

Recruitment

Participants in the lab experiment were recruited from the Stanford university pool on SONA (a web platform). Participants signed up for time slots. Participants in the virtual lab study were recruited from the Stanford and Columbia university pools on SONA . As with any study where participants volunteer to participate, these studies are in principle susceptible to self-selection bias. Although we have no concrete evidence supporting this, it is possible that participants who choose to participate in lab studies are more well-adjusted than other college students, or maybe more social. However, this does not pose a threat to the internal integrity of the study because participants were unaware of the experimental treatment or hypotheses when signing up. Thus, there was no self-selection into conditions. It is possible that the effect size of virtual interaction depends on the sociality of users. However, given that our effects replicated among engineers across multiple countries, self-selection in the lab studies is not a great concern. Participants for the field studies were recruited by their employer to participate in a voluntary workshop. They were paid for their time, but could also opt to not attend. Again, it is possible that participants uninterested in participating in an innovation workshop would not attend to begin with. It is possible that, for people uninterested in innovation, our effects would be weaker. We are not concerned about this because we are interested in the effect of virtual interaction among people who are motivated to perform well, as this reflects most idea generation contexts.

Ethics oversight

We had research clearance from Stanford Human Subjects Committee for all three lab studies and the MTurk survey, and Columbia Human Subjects Committee clearance for the virtual lab study. Participants in all lab studies and Mturk study completed consent forms. Participants in the field studies were not given consent forms because the data was collected by the employer as part of a voluntary workshop they conducted during work hours. We were provided with non-identifiable secondary data, and the manager of the Stanford Human Subjects Committee deemed that no IRB was required for this secondary data source. We received permission from the company to analyze the data given that we would not share the raw data, as these ideas are the intellectual property of the employer.

Note that full information on the approval of the study protocol must also be provided in the manuscript.