Article

# Virtual communication curbs creative idea generation

https://doi.org/10.1038/s41586-022-04643-y

Received: 17 July 2020

Accepted: 14 March 2022

Published online: 27 April 2022

Check for updates

Melanie S. Brucks[1]✉ & Jonathan Levav[2]

COVID-19 accelerated a decade-long shift to remote work by normalizing working from home on a large scale. Indeed, 75% of US employees in a 2021 survey reported a personal preference for working remotely at least one day per week[1], and studies estimate that 20% of US workdays will take place at home after the pandemic ends[2]. Here we examine how this shift away from in-person interaction affects innovation, which relies on collaborative idea generation as the foundation of commercial and scientific progress[3]. In a laboratory study and a field experiment across five countries (in Europe, the Middle East and South Asia), we show that videoconferencing inhibits the production of creative ideas. By contrast, when it comes to selecting which idea to pursue, we find no evidence that videoconferencing groups are less effective (and preliminary evidence that they may be more effective) than in-person groups. Departing from previous theories that focus on how oral and written technologies limit the synchronicity and extent of information exchanged[4–6], we find that our effects are driven by differences in the physical nature of videoconferencing and in-person interactions. Specifically, using eye-gaze and recall measures, as well as latent semantic analysis, we demonstrate that videoconferencing hampers idea generation because it focuses communicators on a screen, which prompts a narrower cognitive focus. Our results suggest that virtual interaction comes with a cognitive cost for creative idea generation.

In the wake of the COVID-19 pandemic, millions of employees were mandated to work from home indefinitely and virtually collaborate using videoconferencing technologies. This unprecedented shift to full-time remote employment demonstrated the viability of virtual work at a large scale, further legitimizing the growing work-from-home movement of the last decade. In a 2021 survey, 75% of US employees reported a personal preference for working from home at least one day a week, and 40% of employees indicated they would quit a job that required full-time in-person work[1]. In response, leading firms across various sectors, including Google, Microsoft, JPMorgan and Amazon, increased the flexibility of their post-pandemic work-from-home policies[7], and research estimates that 20% of all US workdays will be conducted remotely once the pandemic ends[2].

We explore how this shift towards remote work affects essential workplace tasks. In particular, collaborative idea generation is at the heart of scientific and commercial progress[3,8]. From the Greek symposium to Lennon and McCartney, collaborations have provided some of the most important ideas in human history. Until recently, these collaborations have largely required the same physical space because the existing communication technologies (such as letters, email and phone calls) limited the extent of information that is available to communicators and reduced the synchronicity of information exchange (media richness theory, social presence theory, media synchronicity theory[4–6]). However, recent advances in network quality and display resolution have ushered in a synchronous, audio-visual technology—videoconferencing—that conveys many of the same aural and non-verbal information cues as face-to-face interaction. If videoconferencing eventually closes the information gap between virtual and in-person interaction, the question arises whether this new technology could effectively replace in-person collaborative idea generation.

Here we show that, even if video interaction could communicate the same information, there remains an inherent and overlooked physical difference in communicating through video that is not psychologically benign: in-person teams operate in a fully shared physical space, whereas virtual teams inhabit a virtual space that is bounded by the screen in front of each member. Our data suggest that this physical difference in shared space compels virtual communicators to narrow their visual field by concentrating on the screen and filtering out peripheral visual stimuli that are not visible or relevant to their partner. According to previous research that empirically and neurologically links visual and cognitive attention[9–13], as virtual communicators narrow their visual scope to the shared environment of a screen, their cognitive focus narrows in turn. This narrowed focus constrains the associative process underlying idea generation, whereby thoughts ‘branch out’ and activate disparate information that is then combined to form new ideas[14–17]. Yet the narrowed cognitive focus induced by the use of screens in virtual interaction does not hinder all collaborative activities. Specifically, idea generation is typically followed by selecting which idea to pursue, which requires cognitive focus and analytical reasoning[18]. Here we show that virtual interaction uniquely hinders idea

[1]Marketing Division, Columbia Business School, Columbia University, New York, NY, USA. [2]Marketing Division, Stanford Graduate School of Business, Stanford University, Stanford, CA, USA. ✉e-mail: mb4598@columbia.edu

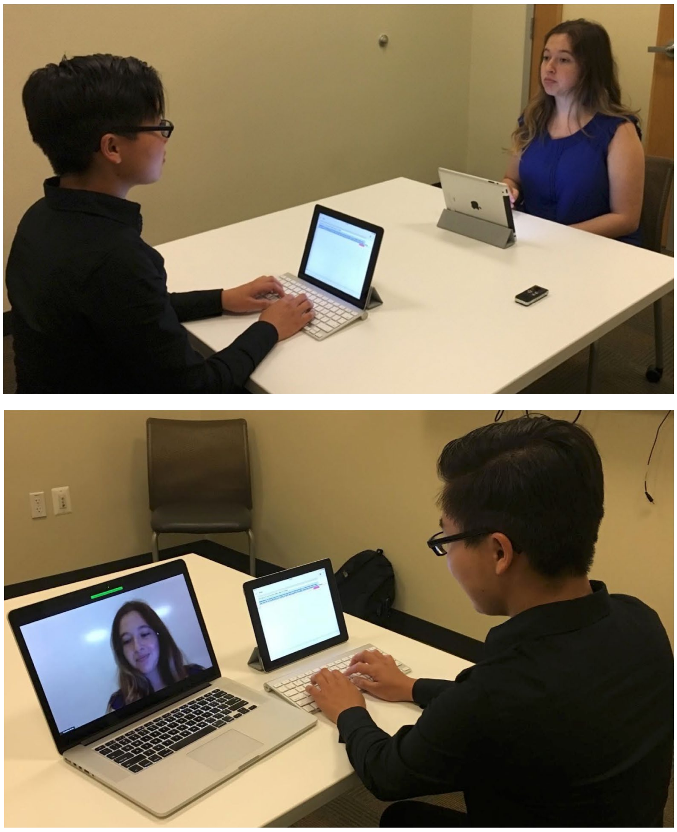


**Fig. 1 | Laboratory experiment set-up.** In the laboratory experiment, we randomly assigned half of the pairs to work together in person and the other half to work together in separate, identical rooms using videoconferencing. The pairs in the virtual condition interacted with a real-time video of their partner's face displayed on a 15-inch retina-display screen with no self-view. The image was taken during the first batch of data collection in the laboratory. Consent was obtained to use these images for publication.

generation—we find that videoconferencing groups generate fewer creative ideas than in-person groups due to narrowed visual focus, but we find no evidence that videoconferencing groups are less effective when it comes to idea selection.

## Laboratory experiment

We first recruited 602 people to participate in an incentive-aligned laboratory study across two batches of data collection (see the 'Laboratory experiment' section of the Methods). The participants were randomly paired, and we instructed each pair to generate creative uses for a product for five minutes and then spend one minute selecting their most creative idea. Pairs were randomly assigned to work together on these tasks either in person or virtually (with their partner displayed by video across from them and the self-view removed; Fig. 1). We assessed ideation performance by counting both the total number of ideas and the subset of creative ideas generated by each pair[14,16,19–23]. We assessed idea selection quality using two different measures: (1) the 'creativity score' of the pair's selected idea[23] and (2) the 'decision error score'—the difference in creativity score between the top scoring idea and the selected idea—where smaller values reflect a better decision[23,24] (see the 'Dependent measures' section of the Methods).

Virtual pairs generated significantly fewer total ideas (mean ($M$) = 14.74, s.d. = 6.23) and creative ideas ($M$ = 6.73, s.d. = 3.27) than in-person pairs (total ideas: $M$ = 16.77, s.d. = 7.27, negative binomial regression, $n$ = 301 pairs, $b$ = 0.13, s.e. = 0.05, $z$ = 2.72, $P$ = 0.007, Cohen's $d$ = 0.30, 95% confidence interval (CI) = 0.07–0.53; creative ideas: $M$ = 7.92, s.d. = 3.40, negative binomial regression, $n$ = 301 pairs, $b$ = 0.16, s.e. = 0.05, $z$ = 3.14, $P$ = 0.002, Cohen's $d$ = 0.36, 95% CI = 0.13–0.58; see Extended Data Table 1 for a summary of all of the analyses, Extended Data Table 2 for results from alternative models and Supplementary Information A for model assumption tests). By contrast, we found indications that virtual interaction might increase decision quality. Virtual pairs selected a significantly higher scoring idea ($M$ = 4.28, s.d. = 0.81) and had a significantly lower decision error score ($M$ = 0.78, s.d. = 0.67) compared with in-person pairs (selected idea: $M$ = 4.08, s.d. = 0.84, linear regression, $n$ = 292 pairs, $b$ = 0.20, s.e. = 0.10, $t_{290}$ = 2.04, $P$ = 0.043, Cohen's $d$ = 0.24, 95% CI = 0.01–0.47; error score: $M$ = 1.01, s.d. = 0.77, linear regression, $n$ = 292 pairs, $b$ = 0.23, s.e. = 0.08, $t_{290}$ = 2.69, $P$ = 0.007, Cohen's $d$ = 0.32, 95% CI = 0.08–0.55; see Supplementary Information B for model assumption tests). However, the effect of modality on decision quality attenuated when controlling for the number of ideas that each pair generated (selected idea: linear regression, $n$ = 292 pairs, $b$ = 0.18, s.e. = 0.10, $t_{289}$ = 1.88, $P$ = 0.062; error score: linear regression, $n$ = 292 pairs, $b$ = 0.20, s.e. = 0.08, $t_{289}$ = 2.40, $P$ = 0.017).

We next examined our hypothesis that virtual communication hampers idea generation because the bounded virtual space shared by pairs narrows visual scope, which in turn narrows cognitive scope. Specifically, in the second batch of data collection, 151 randomly assigned pairs generated creative uses for a product, either inperson or virtually, in a laboratory room containing ten props (five expected, such as folders; and five unexpected, such as a skeleton poster; see the 'Stimulus 2: bubble wrap' section of the Methods; Extended Data Fig. 1). We next captured visual focus in two ways. First, at the end of the study, the participants were asked to individually recall the props in the room and indicate them on a worksheet. Second, we recorded and extracted participants' eye gaze throughout the task (Fig. 2; see the 'Stimulus 2 process measures' section in the Methods).

Supporting our proposition that virtual pairs narrow their visual focus to their shared environment (that is, the screen), virtual pairs spent significantly more time looking directly at their partner ($M_{virtual}$ = 91.4 s, s.d. = 58.3, $M_{in\text{-}person}$ = 51.7 s, s.d. = 52.2, linear mixed-effect regression, $n$ = 270 participants, $b$ = 39.70, s.e. = 6.83, $t_{139}$ = 5.81, $P$ < 0.001, Cohen's $d$ = 0.71, 95% CI = 0.47–0.96), spent significantly less time looking at the surrounding room ($M_{virtual}$ = 32.4 s, s.d. = 34.8, $M_{in\text{-}person}$ = 61.0 s, s.d. = 43.1, linear mixed-effect regression, $n$ = 270 participants, $b$ = 28.75, s.e. = 5.10, $t_{143}$ = 5.64, $P$ < 0.001, Cohen's $d$ = 0.74, 95% CI = 0.49–0.99; Fig. 2) and remembered significantly fewer unexpected props in the surrounding room ($M_{virtual}$ = 1.53, s.d. = 1.38, $M_{in\text{-}person}$ = 1.95, s.d. = 1.38, Poisson mixed-effect regression, $n$ = 302 participants, $b$ = 0.25, s.e. = 0.10, $z$ = 2.47, $P$ = 0.014, Cohen's $d$ = 0.30, 95% CI = 0.08–0.53) than in-person pairs. There was no evidence that the time spent looking at the task differed by modality ($M_{virtual}$ = 176.1 s, s.d. = 64.0, $M_{in\text{-}person}$ = 186.8 s, s.d. = 60.2, linear mixed-effect regression, $n$ = 270 participants, $b$ = 10.65, s.e. = 7.60, $t_{268}$ = 1.40, $P$ = 0.162, Cohen's $d$ = 0.17, 95% CI = −0.07–0.41; see Supplementary Information C for model assumption tests).

Importantly, unexpected prop recall and gaze around the room were both significantly associated with an increased number of creative ideas (room recall: negative binomial regression, $n$ = 151 pairs, $b$ = 0.09, s.e. = 0.03, $z$ = 2.82, $P$ = 0.005; room gaze: negative binomial regression, $n$ = 146 pairs, $b$ = 0.003, s.e. = 0.001, $z$ = 3.14, $P$ = 0.002), and both of these measures independently mediated the effect of modality on idea generation (10,000 nonparametric bootstraps, recall: 95% CI = −0.61 to −0.01; gaze: 95% CI = −1.14 to −0.08; Extended Data Figs. 2 and 3). This combination of analyses converges on the view that virtual communication narrows visual focus, which subsequently hampers idea generation.

These findings provide causal evidence indicating that virtual (versus in-person) interaction hampers idea generation. However, this highly controlled laboratory paradigm may not fully capture the creative process as it unfolds in typical workplaces. Thus, to test the generalizability of the results, we replicated the study in an actual work context in five country sites of a large multinational telecommunications

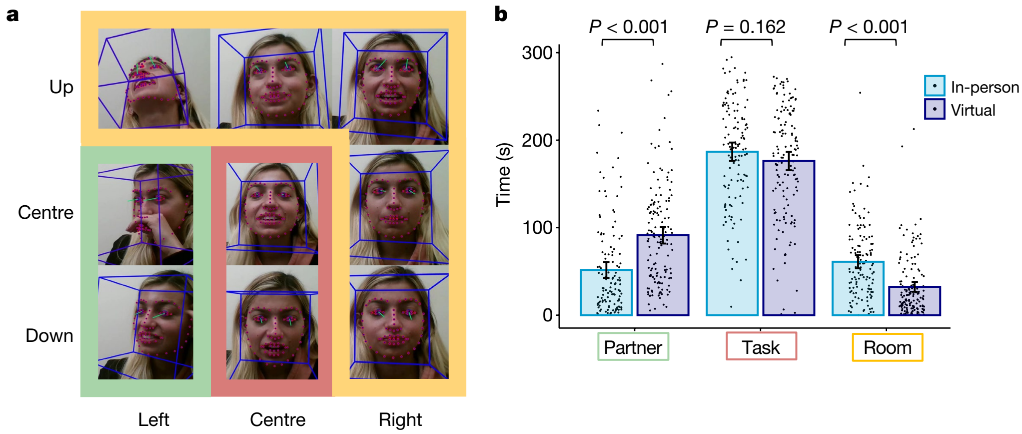



**Fig. 2 | Eye gaze results by modality.** Pairs interacting virtually spent more time looking at their partner ($M_{virtual}$ = 91.4 s, s.d. = 58.3, $M_{in\text{-}person}$ = 51.7 s, s.d. = 52.2, linear mixed-effect regression, $n$ = 270 participants (126 for in-person pairs and 144 for virtual pairs), $b$ = 39.70, s.e. = 6.83, $t_{139}$ = 5.81, $P$ < 0.001, Cohen's $d$ = 0.71, 95% CI = 0.47–0.96) and less time looking at the surrounding room ($M_{virtual}$ = 32.4 s, s.d. = 34.8, $M_{in\text{-}person}$ = 61.0 s, s.d. = 43.1, linear mixed-effect regression, $n$ = 270 participants, $b$ = 28.8, s.e. = 5.10, $t_{143}$ = 5.64, $P$ < 0.001, Cohen's $d$ = 0.74, 95% CI = 0.49–0.99). Importantly, the time spent looking around the room predicted creative idea generation (negative binomial regression, $n$ = 146 pairs (69 in-person pairs and 77 virtual pairs), $b$ = 0.003, s.e. = 0.001, $z$ = 3.14, $P$ = 0.002) and mediated the effect of modality (in-person versus virtual) on idea generation (10,000 nonparametric bootstraps, 95% CI = 1.14–0.08). **a**, Example of how eye gaze during the task was categorized. **b**, Differences in the amount of time (seconds) by modality for looking at one's partner (partner gaze) and looking around the room (room gaze). Data are from the second batch of data collection in the laboratory and are presented as mean ± 95% CIs. All statistical tests were two-sided, and no adjustments were made for multiple comparisons. Consent was obtained to use these images for publication.

infrastructure company. We selected this field setting because it involved domain experts who are highly invested in the outcome, typically know their partners and regularly use virtual-communication technology in their work.

## Field experiment

The company recruited 1,490 engineers to participate in an ideation workshop and randomly assigned the engineers into pairs collaborating either face-to-face or over videoconference (see the 'Field experiment' section of the Methods). The pairs generated product ideas for an hour and then selected and developed one idea to submit as a future product innovation for the company. The engineers who worked on the task virtually ($M$ = 7.43, s.d. = 5.17) generated fewer total ideas than in-person pairs ($M$ = 8.58, s.d. = 6.03, negative binomial mixed-effect regression, $n$ = 745 pairs, $b$ = 0.14, s.e. = 0.05, $z$ = 3.13, $P$ = 0.002, Cohen's $d$ = 0.21, 95% CI = 0.06–0.35). This pattern was replicated at all five sites (Table 1). In three of the workshops ($n$ = 1,238 out of 1,490), engineers served as peer-evaluators and provided external ratings. In these sessions, we found that virtual engineer pairs generated both fewer total ideas ($M$ = 7.42, s.d. = 5.19) and fewer creative ideas ($M$ = 3.83, s.d. = 2.83) compared with in-person teams (total ideas: $M$ = 8.66, s.d. = 6.14, negative binomial mixed-effect regression, $n$ = 619 pairs, $b$ = 0.15, s.e. = 0.05, $z$ = 3.07, $P$ = 0.002, Cohen's $d$ = 0.22, 95% CI = 0.06–0.38; creative ideas: $M$ = 4.32, s.d. = 3.18, negative binomial mixed-effect regression, $n$ = 619 pairs, $b$ = 0.12, s.e. = 0.05, $z$ = 2.15, $P$ = 0.032, Cohen's $d$ = 0.16, 95% CI = 0.01–0.32).

By contrast, we found preliminary evidence that decision quality was positively impacted by virtual interaction. In-person teams had a significantly higher top-scoring idea in their generated idea pool ($M_{virtual}$ = 3.86, s.d. = 0.56, $M_{in\text{-}person}$ = 4.01, s.d. = 0.54, linear mixed-effect regression, $n$ = 619 pairs, $b$ = 0.14, s.e. = 0.04, $t_{608}$ = 3.40, $P$ < 0.001, Cohen's $d$ = 0.27, 95% CI = 0.11–0.43), but the selected idea did not significantly differ in quality by condition ($M_{virtual}$ = 3.05, s.d. = 0.71, $M_{in\text{-}person}$ = 3.04, s.d. = 0.78, linear mixed-effect regression, $n$ = 591 pairs, $b$ = 0.004, s.e. = 0.06, $t_{582}$ = 0.07, $P$ = 0.945, Cohen's $d$ = 0.01, 95% CI = −0.16–0.17). Furthermore, virtual pairs and in-person pairs significantly differ in their decision error score ($M_{virtual}$ = 0.81, s.d. = 0.76, $M_{in\text{-}person}$ = 0.99, s.d. = 0.86, Kruskal–Wallis rank sum, $n$ = 591 pairs, $\chi^2_1$ = 5.30, $P$ = 0.021, Cohen's $d$ = 0.22, 95% CI = 0.05–0.38, but this effect was attenuated when controlling for number of ideas (linear mixed-effect regression, $n$ = 591 pairs, $b$ = 0.11, s.e. = 0.06, $t_{580}$ = 1.81, $P$ = 0.071).

## Alternative explanations

We examined several alternative explanations for the negative effect of virtual interaction on idea generation. A summary of our findings is shown in Extended Data Table 3.

### Incremental ideas

We found that in-person communicators generate a greater number of total ideas and creative ideas compared with virtual pairs. Here, we examine the possibility that the additional ideas that in-person pairs generated could simply be incremental ideas that are topically similar to each other. Specifically, to test the link between virtual communication and associative thinking and whether in-person groups are truly engaging in divergent thinking rather than simply generating ideas 'in the same vein', we used latent semantic analysis to calculate how much each new idea semantically departed from the 'thought stream' of previous ideas[25] in each pair. When ideation is divergent, ideas should depart from preceding ideas. In contrast to the alternative explanation that in-person teams simply generate many similar ideas, we found that, if anything, in-person pairs generated progressively more disconnected ideas over time relative to virtual pairs (Extended Data Fig. 4 and Supplementary Information D). These results are consistent with our proposed process that virtual communication constrains thinking relative to in-person pairs.

### Trust and connection

Previous research found that feelings of connection and trust can facilitate team creativity[26,27]. To examine whether virtual groups experience reduced feelings of connection and whether reduced feelings of connection underlies the negative effect of virtual communication on idea generation, we used three complementary approaches.

First, we examined whether modality affects subjective feelings of connection using data collected through surveys at the end of the laboratory study. Consistent with previous research[28], we found that participants did not report significant differences in feelings of similarity or liking,

**Table 1 | Descriptive statistics**

| | Total number of ideas | | Number of creative ideas | | Decision error score | |
|---|---|---|---|---|---|---|
| | Virtual | In-person | Virtual | In-person | Virtual | In-person |
| Laboratory experiment: stimulus 1 | 13.89 (5.65) | 16.15 (5.51) | 6.17 (3.21) | 7.36 (2.89) | 0.68 (0.65) | 0.92 (0.69) |
| Laboratory experiment: stimulus 2 | 15.57 (6.67) | 17.41 (8.70) | 7.27 (3.26) | 8.49 (3.79) | 0.88 (0.67) | 1.09 (.82) |
| Field: Portugal | 7.47 (5.11) | 8.12 (5.65) | | | | |
| Field: Israel | 7.17 (5.27) | 8.42 (4.94) | | | | |
| Field: Finland | 8.56 (4.59) | 10.19 (5.91) | 4.11 (2.49) | 5.00 (3.17) | 0.91 (0.79) | 1.44 (.88) |
| Field: Hungary | 6.83 (4.95) | 8.15 (5.82) | 3.26 (2.58) | 3.86 (2.85) | 0.72 (0.76) | 0.95 (0.85) |
| Field: India | 7.65 (5.42) | 8.75 (6.37) | 4.18 (2.99) | 4.51 (3.37) | 0.85 (0.76) | 0.93 (0.84) |

Descriptive statistics for the number of ideas generated, number of creative ideas generated, and decision error score by modality. We calculated the decision error score by taking the difference in creativity score between the top scoring idea and the selected idea—smaller values reflect a better decision. Data are mean (s.d). The laboratory study collected data from 301 pairs, 150 pairs in the first batch of data collection and 151 pairs in the second batch of data collection. The field study was conducted with a total 1,490 engineers, which amounted to 745 pairs (Portugal, 105 pairs; Israel, 18 pairs; Finland, 54 pairs; Hungary, 230 pairs; India, 338 pairs).

or in perceptions of how 'in sync' they were as a team by modality. Supporting these self-reported feelings, in an economic trust game, virtual and in-person pairs did not significantly differ in the amount of money they entrusted to their partner, although the amount of money entrusted to their partner was positively correlated with the number of creative ideas that the pair generated (details of the methods and analyses in this paragraph are provided in Supplementary Information E).

Second, we examined whether modality affects social behaviours (both verbal and nonverbal) by extracting behavioural data from unobtrusive recordings of laboratory participants during the idea-generation task. Specifically, we quantified social behaviour using two methods: (1) judges blinded to the condition and hypotheses watched video clips of the participants and scored the extent to which they observed 32 social behaviours; and (2) we transcribed each pair's conversation and ran the transcripts through a linguistic analysis database[29]. Across these analyses, we found that that virtual and in-person pairs significantly differed in only 4 out of 32 observer-rated behaviours and the word usage of only 3 out of 80 social and cognitive linguistic categories in the database. Furthermore, we found that controlling for these differences did not significantly attenuate the negative effect of virtual interaction on idea generation (see Extended Data Table 4 for a summary and Supplementary Information F and G for details of the methods and analyses in this paragraph).

Finally, we examined whether modality affects mimicry, a subconscious indicator of connection[30]. Specifically, we assessed the extent to which pairs exhibited linguistic mimicry[31] using each pair's transcripts and the extent to which pairs exhibited facial mimicry[30,32] using facial expressions extracted from the videos of their interactions[33]. We found that in-person and virtual pairs did not significantly differ in the extent to which they exhibited either form of mimicry (see Supplementary Information H for details of the methods and analyses in this paragraph).

These results demonstrate how similar video interactions can be to in-person communication. Across three complementary approaches (subjective feelings of closeness, verbal and non-verbal behaviours, and mimicry), we found little evidence that communication modality affects social connection. Furthermore, the significant negative effect of virtual interaction on idea generation holds when controlling for these measures. Thus, it seems improbable that differences in social connection or social behaviour by interaction modality are a main contributor to the results that we report.

### Conversation coordination

Although video and in-person interaction contains many of the same informational cues, one important distinction between these modalities is the ability to engage in eye contact. When two individuals look at each other's eyes on the screen, it appears to neither partner that the other is looking into their eyes, which could affect communication coordination[34]. Indeed, previous research found that virtual pairs can experience difficulty in determining who should speak next and when[35] and, in our studies, virtual pairs reported struggling more with communication coordination. We used three complementary metrics from the laboratory study transcripts to examine whether communication coordination friction contributes to our effect: the number of words spoken, the number of times the transcriber observed 'crosstalk' during the interaction (which reflects when two people speak over each other) and the number of speaker switches (back-and-forth) that each pair exhibited. The number of words spoken did not significantly differ by modality, but virtual groups engaged in significantly fewer speaker switches compared with in-person groups and significantly less crosstalk. However, controlling for these measures did not significantly attenuate the effect of modality on number of creative ideas generated (see Supplementary Information I for details of the methods and analyses in this paragraph). Together, although communication coordination is altered by the modality of interaction, it does not appear to fully explain the effect of virtual interaction on idea generation.

### Interpersonal processes

In addition to social connection and conversation coordination, previous research has identified a range of interpersonal processes that can affect group idea generation: fear of evaluation (and resulting self-censorship), dominance, social facilitation, social loafing, social sensitivity, perceptions of performance and production blocking. We examined whether interaction modality alters any of these processes during idea generation and whether controlling for these processes meaningfully attenuated our documented negative effect of virtual collaboration on ideation (see Supplementary Information J for a relevant literature and details of the methods and analyses). In these supplementary analyses, we found that our observed effect was robust to each of these alternative explanations.

## Policy implications

Our results suggest that there is a unique cognitive advantage to in-person collaboration, which could inform the design of remote work policies. However, when determining whether or not to use virtual teams, many additional factors necessarily enter the calculus, such as the cost of commute and real estate, the potential to expand the talent pool, the value of serendipitous encounters[36], and the difficulties in managing time zone and regional cultural differences[37]. Although some of these factors are intangible and more difficult to quantify, there are concrete and immediate economic advantages to virtual interaction (such as a reduced need for physical space, reduced salary for employees in areas with a lower cost of living and reduced business travel expenses). To capture the best of both worlds, many workplaces are planning to or currently combine in-person and virtual interaction.

Indeed, a 2021 survey suggests that American employees will work from home around 20% of the time after the pandemic[2]. Our results indicate that, in these hybrid setups, it might make sense to prioritize creative idea generation during in-person meetings. However, it is important to caution that our results document only the cognitive cost of virtual interaction. When it comes to deciding the extent to which a firm should use virtual teams, a more comprehensive analysis factoring in other industry and context-specific costs that the firm might face is needed. We leave this important issue to future research.

## Extensions and generalizability

Our studies examined the effect of virtual versus in-person interaction in the context of randomly assigned pairs (a context relevant to practice; Supplementary Information K). Here, we explore whether the negative effect of virtual interaction on idea generation generalizes to more established and/or larger teams. To investigate this, we first examined the relevance to established teams by testing whether familiarity moderated the effect of modality on ideation in our field data. We found no evidence that the negative effect of virtual interaction changed depending on level of familiarity between members of the pair (Supplementary Information L). This hints that the effect of virtual interaction on idea generation might extend to established teams. Second, we examined the implications of our effect on larger teams by manipulating the virtual group size in an online experiment. Replicating previous findings on the counterproductivity of larger in-person ideation teams[21], virtual pairs outperformed larger virtual groups (see Extended Data Fig. 5 and Extended Data Table 5 for summaries of the set-up and our findings, and Supplementary Information M for details of the methods and analysis). Putting our findings together with previous findings, research suggests that in-person pairs outperform larger in-person groups and virtual pairs, and virtual pairs outperform larger virtual groups. Thus, our substantive recommendation is, cost permitting, to generate ideas in pairs and in person.

In our laboratory study, we used 15.6-inch MacBook Pro retina display laptops because, at the time of data collection (2016 to 2021), laptops were the most common hardware used in videoconferencing, and 15.6 inch was the most prevalent screen size offered in the market (Supplementary Information N). We examined whether our effect would generalize to larger screen sizes in the online study described above. Specifically, we leveraged participants' natural variance in screen size to investigate the role of screen size in virtual idea generation. We found no evidence that screen size is associated with idea generation performance in virtual pairs (even when controlling for income, comfort with videoconferencing and time spent on the computer; Supplementary Information M). This suggests that, at least in the current range available on the market, larger screen sizes would not ameliorate the negative effect of virtual interaction on ideation.

Finally, our empirical context involved both novices (college undergraduates) and experts (engineering teams) and examined two types of creativity tasks: a low-complexity task in the laboratory (generating alternative uses for a product) and a high-complexity task in the field (identifying problems that customers might have as well as generating solutions the firm could offer). Although it is reassuring that we demonstrate our effects with two different participant pools that vary in their creativity training, task type and domain expertise, our contexts are representative of only a subset of innovation teams. Future research is needed to examine the moderating factor of group heterogeneity (Supplementary Information O and P) and extensions to other creative industries (Supplementary Information Q).

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04643-y.


1. Barrero, J. M., Bloom, N. & Davis, S. J. Don't force people to come back to the office full time. Harvard Business Review (24 August 2021).
2. Barrero, J. M., Bloom, N. & Davis, S. *Why Working from Home Will Stick*. NBER Working paper (2021); http://www.nber.org/papers/w28731.pdf.
3. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
4. Daft, R. L. & Lengel, R. H. Organizational information requirements, media richness and structural design. *Manage. Sci.* **32**, 554–571 (1986).
5. Short, J., Williams, E. & Christie, B. *The Social Psychology of Telecommunications* (John Wiley & Sons, 1976).
6. Dennis, A. R., Fuller, R. M. & Valacich, J. S. Media, tasks, and communication processes: a theory of media synchronicity. *MIS Q.* **32**, 575–600 (2008).
7. Kelly, J. Here are the companies leading the work-from-home revolution. Forbes (24 May 2020).
8. Salter, A. & Gann, D. Sources of ideas for innovation in engineering design. *Res. Pol.* **32**, 1309–1324 (2003).
9. Anderson, M. C. & Spellman, B. A. On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychol. Rev.* **102**, 68–100 (1995).
10. Derryberry, D. & Tucker, D. M. In *The Heart's Eye: Emotional Influences in Perception and Attention* (eds Niedenthal, P. M. & Kitayama, S.) 167–196 (Academic, 1994).
11. Posner, M. I. *Cognitive Neuroscience of Attention* (Guilford Press, 2011).
12. Rowe, G., Hirsh, J. B. & Anderson, A. K. Positive affect increases the breadth of attentional selection. *Proc. Natl Acad. Sci. USA* **104**, 383–388 (2007).
13. Friedman, R. S., Fishbach, A., Förster, J. & Werth, L. Attentional priming effects on creativity. *Creativity Res. J.* **15**, 277–286 (2003).
14. Mehta, R. & Zhu, R. J. Blue or red? Exploring the effect of color on cognitive task performances. *Science* **323**, 1226–1229 (2009).
15. Mednick, S. The associative basis of the creative process. *Psychol. Rev.* **69**, 220–232 (1962).
16. Nijstad, B. A. & Stroebe, W. How the group affects the mind: a cognitive model of idea generation in groups. *Pers. Soc. Psychol. Rev.* **10**, 186–213 (2006).
17. Jung, R. E., Mead, B. S., Carrasco, J. & Flores, R. A. The structure of creative cognition in the human brain. *Front. Hum. Neurosci.* **7**, 1–13 (2013).
18. Simon, H. A. A behavioral model of rational choice. *Q. J. Econ.* **69**, 99–118 (1955).
19. Diehl, M. & Stroebe, W. Productivity loss in brainstorming groups: toward the solution of a riddle. *J. Pers. Soc. Psychol.* **3**, 497–509 (1987).
20. Osborne, A. F. *Applied Imagination* (Scribner, 1957).
21. Mullen, B., Johnson, C. & Salas, E. Productivity loss in brainstorming groups: a meta-analytic integration. *Basic Appl. Soc. Psychol.* **12**, 3–23 (1991).
22. Luo, L. & Toubia, O. Improving online idea generation platforms and customizing the task structure on the basis of consumers' domain-specific knowledge. *J. Mark.* **79**, 100–114 (2015).
23. Keum, D. D. & See, K. E. The influence of hierarchy on idea generation and selection in the innovation process. *Organ. Sci.* **28**, 653–669 (2017).
24. Berg, J. M. Balancing on the creative highwire: forecasting the success of novel ideas in organizations. *Admin. Sci. Q.* **61**, 433–468 (2016).
25. Gray, K. et al. "Forward Flow": a new measure to quantify free thought and predict creativity. *Am. Psychol.* **74**, 539–55416 (2019).
26. West, M. A. The social psychology of innovation in groups. in *Innovation and Creativity at Work* (eds West, M. A. & Farr, J. L.) 309–322 (John Wiley & Sons, 1990).
27. Nemiro, J. E. Connection in creative virtual teams. *J. Behav. Appl. Manage.* **2**, 93–115 (2016).
28. Sprecher, S. Initial interactions online-text, online-audio, online-video, or face-to-face: effects of modality on liking, closeness, and other interpersonal outcomes. *Comput. Hum. Behav.* **31**, 190–197 (2014).
29. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
30. Chartrand, T. L. & van Baaren, R. Human Mimicry. In *Advances in Experimental Social Psychology* Vol. 41 (ed. Zanna, M. P.) 219–274 (Academic Press, 2009).
31. Ireland, M. E. et al. Language style matching predicts relationship initiation and stability. *Psychol. Sci.* **22**, 39–44 (2011).
32. Kulesza, W. M. et al. The face of the chameleon: the experience of facial mimicry for the mimicker and the mimickee. *J. Soc. Psychol.* **155**, 590–604 (2015).
33. Amos, B., Ludwiczuk, B. & Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6.2 (2016).
34. Jokinen, K., Nishida, M. & Yamamoto, S. On eye-gaze and turn-taking. In *Proc. 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction* 118–123 (ACM Press, 2010); https://doi.org/10.1145/2002333.2002352.
35. Boland, J. E., Fonseca, P., Mermelstein, I. & Williamson, M. Zoom disrupts the rhythm of conversation. J. Exp. Psychol. https://doi.org/10.1037/xge0001150 (2021).
36. Hinds, P. J. & Mortensen, M. Understanding conflict in geographically distributed teams: the moderating effects of shared identity, shared context, and spontaneous communication. *Organ. Sci.* **16**, 290–307 (2005).
37. Mortensen, M. & Hinds, P. J. Conflict and shared identity in geographically distributed teams. *Int. J. Confl. Manage.* **12**, 212–238 (2001).


**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

### Laboratory experiment

Following the methodological recommendation[38] of "generaliz[ing] across stimuli by replicating the study across different stimuli within a single experiment", we collected our laboratory data with stimulus replicates in two batches. Where possible, we combined the data from the two batches to increase statistical power. When the two batches of data were combined, our power to detect a difference in conditions at our effect size was 89%. Below, we outline the methods for each stimulus batch.

### Stimulus 1: frisbee

**Procedure.** Three hundred participants (202 female, 95 male, $M_{age}$ = 26.1; s.d.$_{age}$ = 8.61; three participants did not complete the survey and are therefore missing demographic information) from a university student and staff pool in the United States participated in the study in exchange for US$10. We posted timeslots in an online research portal that allowed each participant to enroll anonymously into a pair. The participants provided consent before beginning the study. This study was approved by the Stanford University Human Subjects Ethics Board (protocol 35916). The laboratory study was conducted by university research assistants blinded to the hypothesis who were not present during the group interaction.

On arrival, the pairs were informed that their first task was to generate creative alternative uses for a Frisbee and that their second task was to select their most creative idea. These tasks were incentive-aligned: each creative idea that was generated (as scored by outside judges) earned the pair one raffle ticket for a US$200 raffle, and selecting a creative idea earned the pair five additional raffle tickets. Half of the teams ($n$ = 75) learned that they would be working together on the task in the same room, whereas the other half ($n$ = 75) were told that they would be working in separate rooms and communicating using video technology (WebEx, v.36.6–36.9). Groups were assigned in an alternating order, such that the first group was in-person, the second group was virtual and so on. This ensured an equal and unbiased recruitment of each condition.

Before being moved to the task room(s), one participant was randomly selected to be the typist (that is, to record the ideas during the idea-generation stage and indicate the selected idea in the idea-selection stage for the pair) by drawing a piece of paper from a mug. In both communication modalities, each team member had an iPad with a blank Google sheet open (accessed in 2016). The typist had a wireless keyboard and editing capabilities, whereas the other team member could only view the ideas on their iPad. Thus, only the typist could record the generated ideas and select the pair's top idea, but both members had equal information about the team's performance (that is, the generated ideas and the selected idea). In-person pairs sat at a table across from each other. Virtual pairs sat at identical tables in separate rooms with their partner displayed on video across from them. The video display was a full-screen video stream of only their partner (the video of the self was not displayed) on a 15-inch retina-display MacBook Pro.

Each pair generated ideas for 5 min and spent 1 min selecting their most creative idea. They indicated their top creative idea by putting an asterisk next to the idea on the Google sheet. Nine groups did not indicate their top idea on the Google sheet; in the second batch of data collection, we used an online survey that required a response to prevent this issue. Finally, as an exploratory measure, each pair was given 5 min to evaluate each of their ideas on a seven-point scale (1 (least creative) to 7 (most creative)).

Once pairs completed both the idea-generation and the idea-selection task, each team member individually completed a survey on Qualtrics (accessed in 2016) in a separate room.

### Stimulus 2: bubble wrap

**Procedure.** Participants (334) from a university student and staff pool in the United States participated in the study in exchange for US$15. We also recruited 18 participants from Craigslist in an effort to accelerate data collection. However, the students reported feeling uncomfortable, and idea generation performance dropped substantially with student–craigslist pairs, so we removed these pairs from the analysis. Our final participant list did not overlap with participants in the first batch of data collection in the laboratory. The participants provided consent before beginning the study. This study was approved by the Stanford University Human Subjects Ethics Board (protocol 35916). The laboratory study was conducted by university research assistants blinded to the hypothesis who were not present during the group interaction.

We a priori excluded any pairs who experienced technical difficulties (such as screen share issues, audio feedback or dropped video calls) and aimed to collect 150 pairs in total. Our final sample consisted of 302 participants (177 females, 119 males, 2 non-binary, $M_{age}$ = 23.5, s.d.$_{age}$ = 7.09; we are missing demographic and survey data from four of the participants). Mimicking the design of the first batch of data collection in the laboratory, pairs generated uses for bubble wrap for 5 min and then spent 1 min selecting their most creative idea. As before, half of the teams learned that they would be working together on the task in the same room ($n$ = 74), whereas the other half ($n$ = 77) were told that they would be working in separate (but identical) rooms and communicating using video technology (Zoom v.3.2). The groups were assigned in an alternating order, such that the first group was in-person, the second group was virtual and so on. This ensured an equal and unbiased recruitment of each condition. Again, one partner was randomly assigned to be the typist. The tasks were incentivized using the same structure as the first batch of collection.

For in-person pairs, each participant had a 15-inch task computer directly in front of them with their partner across from them and situated to their right. For virtual pairs, each participant had two 15-inch computers: a task computer directly in front of them and computer displaying their partner's face to their right (again, self-view was hidden). This set-up enabled us to unobtrusively measure gaze by using the task computer to record each participant's face during the interaction: in both conditions, the task was directly in front of each participant and the partner was to each participant's right.

In contrast to the first batch of data collection, we used Qualtrics (accessed in 2018) to collect task data. Pairs first generated alternative uses for bubble wrap. After 5 min, the page automatically advanced. We next asked each pair to select their most creative idea and defined a creative idea as both novel (that is, different from the normal uses of bubble wrap) and functional (that is, useful and easy to implement). The pair had exactly 1 min to select their most creative idea. After 1 min, the page automatically advanced. If the pair still had not selected their top idea, the survey returned the selection page and marked that the team went over time. Virtual and in-person pairs did not significantly differ in the percentage of teams that went over time (that is, took longer than a minute); 17.6% of in-person pairs and 16.9% of virtual pairs went over time (Pearson's $\chi^2_1$ = 0.001, $P$ = 0.926). Finally, as exploratory measures, each pair (1) selected an idea from another idea set and then (2) evaluated how novel and functional their selected idea was on a seven-point scale.

Importantly, in both conditions, the task rooms were populated with ten props: five expected props (that is, props consistent with a behavioural laboratory schema (a filing cabinet, folders, a cardboard box, a speaker and a pencil box)) and five unexpected props (a skeleton poster, a large house plant, a bowl of lemons, blue dishes and yoga ball boxes; Extended Data Fig. 1, inspired by ref. [39]). Immediately after the task, we moved the participants into a new room, separated them and asked the participants to individually recreate the task room on a sheet of paper[39].

After the room recall, to measure social connection, each participant responded to an incentive-aligned trust game[40]. Specifically, each participant read the following instructions: "Out of the 150 groups in this study, 15 groups will be randomly selected to win $10. This is a REAL bonus opportunity. Out of the $10, you get the choose how much to share with your partner in the study. The amount of money you give to your partner will quadruple, and then your partner can choose how much (if any) of that money they will share back with you."

The participants then selected how much money they would entrust to their partner in US$1 increments, between US$0 and US$10. Finally, the participants then completed a survey with exploratory measures.

### Dependent measures

**Measure of idea generation performance.** Researchers conducting the analyses were not blinded to the hypothesis and all data were analysed using R (v.4.0.1). We first computed total idea count by summing the total number of ideas generated by each pair. Then, for the key dependent measure of creative ideas, we followed the consensual assessment technique[41] and had two undergraduate judges (from the same population and blind to condition and hypothesis) evaluate each idea on the basis of novelty. Specifically, each undergraduate judge was recruited by the university's behavioural laboratory to help code data from a study. Each judge was given an excel sheet with all of the ideas generated by all of the participants in a randomized order and was asked to evaluate each idea for novelty on a scale of 1 (not at all original/innovative/creative) to 7 (very original/innovative/creative) in one column of the excel sheet and to evaluate each idea for value on a scale of 1 (not at all useful/effective/implementable) to 7 (very useful/effective/implementable) in an adjacent column. Anchors were adopted from ref. [42].

Judges demonstrated satisfactory agreeability (stimulus 1: $\alpha_{\text{novelty}} = 0.64$, $\alpha_{\text{value}} = 0.68$, stimulus 2: $\alpha_{\text{novelty}} = 0.75$, $\alpha_{\text{value}} = 0.67$) on the basis of intraclass correlation criteria delineated previously[43]. The scores were averaged to produce one creativity score for each idea. We computed the key measure of creative idea count by summing the number of ideas that each pair generated that surpassed the average creativity score of the study (that is, the grand mean of the whole study for each stimulus across the two conditions). Information about average creativity is provided in Supplementary Information R.

**Measure of selection performance.** We followed previous research and calculated idea selection using two different methods[23,24]. First, we examined whether the creativity score of the idea selected by each pair differed by communication modality (both with and without controlling for the creativity score of the top idea). Second, we calculated the difference between the creativity score of the top idea and the creativity score of the selected idea. A score of 0 indicates that they selected their top-scoring idea, and a higher score reflects a poorer decision.

### Stimulus 2 process measures

**Room recall.** The room contained five expected props and five unexpected props. If virtual participants are more visually focused, they should recall fewer props and, specifically, the unexpected props that cannot be guessed using the schema of a typical behavioural laboratory. To test this, we counted the number of total props (out of the ten) and unexpected props (out of five) that participants drew and labelled when sketching the room from memory. We did not include other objects in the room (such as the computer and door) in our count.

**Eye gaze.** We used OpenFace (v.2.2.0), an opensource software package, to automatically extract and quantify eye gaze angles using the recording of each participant taken from their task computer[34]. From there, we had at least two independent coders (blinded to the hypothesis and condition) look at video frames of eye gaze angles extracted from the software and indicate the idiosyncratic threshold at which each participant's eye gaze shifted horizontally (from left to centre, and centre to right, $\alpha = 0.98$) and vertically (up to centre, and centre to down, $\alpha = 0.85$). Out of 302 participants, 275 videos of participants yielded usable gaze data. 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.

Using these thresholds, we calculated how often each participant looked at their partner, the task and the surrounding room. To repeat, the recording came from the task computer, and the partner was always situated to the participant's right (or from the perspective of a person viewing the video, to the left). As human coders marked the thresholds (blind to the hypothesis and condition), we report the categorizations from the perspective of an observer of the video. Specifically, looking either (1) horizontally to the left and vertically centre or (2) horizontally to the left and vertically down was categorized as 'partner gaze'; looking either (1) horizontally centre and vertically centre or (2) horizontally centre and vertically down was categorized as 'task gaze'; and the remaining area was categorized as 'room gaze', which encompassed looking (1) horizontally left and vertically upward, (2) horizontally centre and vertically upward, (3) horizontally right and vertically upward, (4) horizontally right and vertically centre, and (5) horizontally right and vertically down (Fig. 2; consent was obtained to use these images for publication). We chose this unobtrusive methodology instead of more cumbersome eye-tracking hardware to maintain organic interactions—wearing strange headgear could make participants consciously aware of their eye gaze or change the natural dynamic of conversation.

We excluded six videos that were less than 290 s long. The effects do not change in significance when these videos are included in the analyses. With these excluded videos, as before, virtual groups spent significantly more time looking at their partner ($M_{\text{virtual}} = 90.6$ s, s.d. = 58.3, $M_{\text{in-person}} = 52.6$ s, s.d. = 54.3, linear mixed-effect regression, $n = 276$ participants, $b = 38.00$, s.e. = 6.95, $t_{139} = 5.46$, $P < 0.001$, Cohen's $d = 0.68$, 95% CI = 0.43–0.92) and spent significantly less time looking at the surrounding room ($M_{\text{virtual}} = 32.4$ s, s.d. = 34.6, $M_{\text{in-person}} = 60.9$ s, s.d. = 43.7, linear mixed-effect regression, $n = 276$ participants, $b = 28.44$, s.e. = 4.96, $t_{145} = 5.74$, $P < 0.001$, Cohen's $d = 0.73$, 95% CI = 0.48–0.98; Fig. 2). There was again no evidence that time spent looking at the task differed by modality ($M_{\text{virtual}} = 176$ s, s.d. = 63.6, $M_{\text{in-person}} = 184$ s, s.d. = 63.0, linear mixed-effect regression, $n = 276$ participants, $b = 7.39$, s.e. = 7.63, $t_{274} = 0.97$, $P = 0.334$, Cohen's $d = 0.12$, 95% CI = −0.12–0.35). Importantly, gaze around the room was significantly associated with an increased number of creative ideas (negative binomial regression, $n = 146$ pairs, $b = 0.003$, s.e. = 0.001, $z = 3.10$, $P = 0.002$). Furthermore, gaze around the room mediated the effect of modality on idea generation (5,000 nonparametric bootstraps, 95% CI = 0.05 to 1.15).

## Field experiment

### Procedure

Engineer teams (837) participated in an ideation workshop under the aegis of an 'intrapreneurship' program. The workshop was conducted by the employer at company offices in Finland, Hungary, India, Israel and Portugal as part of a voluntary workshop during work hours. The employer shared non-identifiable secondary data from the workshop with us and, as a result, the Stanford Human Subjects Committee deemed that no IRB or consent was required. The company gave us permission to publish the results from our analysis, contingent on their intellectual property from the workshops (the generated ideas) remaining confidential. At each site, typically two sessions were run each day, for a total of 26 sessions across sites, and each session was made up of an organizational team (average session size, 32 teams; range of session size, 8 teams to 56 teams). Research assistants (who were not blinded to the hypothesis) ran the workshop and were present during group interaction to answer any questions and fix technical problems.

In the workshop, engineers were first given a 60 min introduction to the intrapreneurship program. In Hungary, for the first six sessions, there was an additional 60 min lecture about creativity and idea selection before the workshop. After observing fatigue, the order was switched for the remaining two sessions. Engineers then learned that the workshop involved (1) generating new product ideas for the company and then (2) selecting and developing one idea to submit to the intrapreneurship program. Engineers were instructed that the ideas should not be limited to current products offered by the company and to consider ideas that could become products two years from now.

The engineers were randomly assigned into pairs, with one engineer randomly assigned as the typist, and the pairs were randomly assigned into condition (in-person versus virtual). The in-person and virtual conditions were split into two different rooms and then issued the manipulation. Pairs in the virtual condition ($n = 434$) learned that, because they worked at a global company, the company was testing different collaboration tools and the engineers would therefore be doing this workshop virtually over Webex (a communication technology almost all the engineers were familiar with and often used to collaborate with their remote teams). The following Webex versions were used in each workshop: Portugal, 38.3; Finland and Israel, 38.6; Hungary, 39.3; India, 40.1. The typists remained in one room and their partners were brought to another room. All of the pairs established their video connection before the timed idea generation session began. Pairs in the in-person condition ($n = 403$) were instructed to space out and find a place in the room to sit together before beginning the idea generation task.

In Finland, Israel and Portugal, the pairs had 1 h to generate ideas. In Hungary and India, the pairs had 45 min to generate ideas to allow for time at the end to peer-evaluate each other's ideas. After idea generation, the pairs selected one of their ideas to further develop and submit to the intrapreneurship program. In Finland, Israel and Portugal, the firm used a third-party platform for idea collection that didn't allow for idea rating (Innostreams, beta release, v.0.8). However, in Hungary and India, the employer designed a workshop platform on Qualtrics (accessed in 2019 and 2020) and implemented a peer-evaluation system at the end of each workshop session. Specifically, each engineer was assigned 8 to 12 ideas from other pairs in the idea generation session. Furthermore, to increase power, the company also sent out a Qualtrics survey (accessed in 2021 and 2022) to their engineers to score the ideas collected in Finland. Following the same procedure as in the workshop, engineers were randomly assigned ten ideas to score. Given low participation rates, the employer allowed engineers to submit up to three scores. We analysed the first three submissions from each engineer. A small subset of engineers participated more than ten times ($n = 6$) and were excluded from analyses. Participation was voluntary, and each submission won a raffle ticket into a raffle for an iPad.

To score the ideas, the engineers individually rated each idea on four dimensions identified by the intrapreneurship program: (1) this idea identifies a real pain point that is significant to customers; (2) the idea is potentially disruptive and game changing; (3) this idea has a unique value proposition; and (4) the solution is realistic. Idea scores were calculated by averaging these four dimensions for each idea. To capture idea selection, we calculated a 'decision error score' by subtracting the score of the highest-rated idea from the score of the selected idea.

#### Exclusions

One site in Poland was predetermined to be excluded. Unlike the others, this workshop was conducted by the company outside of company offices, in a hotel meeting space. Perhaps for this reason, participants exhibited rampant non-compliance, including a notable preoccupation with the hotel catering's coffee and cookie station. The non-compliance was exacerbated in the in-person condition, as many engineers brought coffee and snacks back to their workspace and socialized with other teams; the headsets in the virtual condition prohibited the same extent of socialization. Importantly, the results do not change in significance if we include this sixth site ($M_{\mathrm{virtual}} = 7.38$, s.d. = 5.18, $M_{\mathrm{in\text{-}person}} = 8.15$, s.d. = 5.78, negative binomial mixed-effect regression, $n = 904$ pairs, $b = 0.09$, s.e. = 0.04, $z = 2.24$, $P = 0.025$, Cohen's $d = 0.14$, 95% CI = 0.01–0.27).

Across the remaining sites, we a priori dropped 32 pairs due to technical difficulties, 21 pairs due to non-compliance and 18 teams because they arrived late or left early. We also excluded 5 teams who had participated previously and 15 three-person teams (when the session size contained an odd number of people). Thus, out of 837 teams, 745 pairs (1,490 engineers) constituted our final sample. 365 pairs were assigned to interact in person, and 380 groups were assigned to interact virtually. There are slightly more virtual groups because, after a few workshops, the company began to over-assign Webex groups by creating more odd labels than even labels to pass out to compensate for a few Webex groups each session being dropped due to technical difficulties.

Furthermore, for idea selection, we a priori dropped teams that reselected their idea (that is, selected an idea, worked on it and then changed their mind, $n = 14$ pairs), and we are missing data from groups that did not select their idea ($n = 1$ pair), did not submit their ideas in time to get scored ($n = 3$ pairs), and teams whose ideas were not scored due to attrition during the idea evaluation survey ($n = 13$ pairs). Thus, out of 619 pairs in the final two sites, 305 in-person pairs and 317 virtual pairs constituted our final sample for idea selection. The results do not differ in significance if we include all of the excluded teams (idea generation: $M_{\mathrm{virtual}} = 7.44$, s.d. = 5.51, $M_{\mathrm{in\text{-}person}} = 8.33$, s.d. = 5.99, negative binomial mixed-effect regression, $n = 822$ pairs, $b = 0.11$, s.e. = 0.05, $z = 2.52$, $P = 0.012$, Cohen's $d = 0.16$, 95% CI = 0.02–0.29; idea selection: $M_{\mathrm{virtual}} = 0.79$, s.d. = 0.76, $M_{\mathrm{in\text{-}person}} = 0.92$, s.d. = 0.84, Kruskal–Wallis rank sum, $n = 656$ pairs, $\chi^2_1 = 5.46$, $P = 0.019$, Cohen's $d = 0.21$, 95% CI = 0.05–0.36). Note that we still excluded three-person teams because they deviate from our context.

Researchers conducting the analyses were not blinded to the hypothesis and all data were analysed using R (v.4.0.1).

#### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The data (raw and cleaned) collected by the research team and reported in this Article and its Supplementary Information are available on Research Box (https://researchbox.org/282), except for the video, audio recordings and transcripts of participants, because we do not have permission to share the participants' voices, faces or conversations. The cleaned summary data for the field studies are available in the same Research Box, but the raw data must be kept confidential, as these data are the intellectual property of the company. The Linguistic Analysis database is available online (https://liwc.wpengine.com/). Extended Data Tables 1–5 and Extended Data Figs. 2 and 3 are summary tables and figures, and the raw data associated with these tables are on Research Box (https://researchbox.org/282). Source data are provided with this paper.

## Code availability

All custom code used to clean and analyse the data is available at Research Box (https://researchbox.org/568). The Linguistic Analysis database is available online (https://liwc.wpengine.com/). OpenFace is available at GitHub (https://github.com/TadasBaltrusaitis/OpenFace).

38. Kenny, D. A. Quantitative Methods for social psychology. In *Handbook of Social Psychology* Vol. 1 (eds Lindzey, G. & Aronson, E.) 487–508 (Random House, 1985).
39. Brewer, W. F. & Treyens, J. C. Role of schemata in memory for places. *Cogn. Psychol.* **13**, 207–230 (1981).
40. King-Casas, B. Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78–83 (2005).
41. Amabile, M. Social psychology of creativity: a consensual assessment technique. *J. Pers. Soc. Psychol.* **43**, 997–1013 (1982).
42. Moreau, C. P. & Dahl, D. W. Designing the solution: the impact of constraints on consumers' creativity. *J. Consum. Res.* **32**, 13–22 (2005).

43. Cicchetti, D. V. & Sparrow, S. A. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am. J. Ment. Def.* **86**, 127–137 (1981).

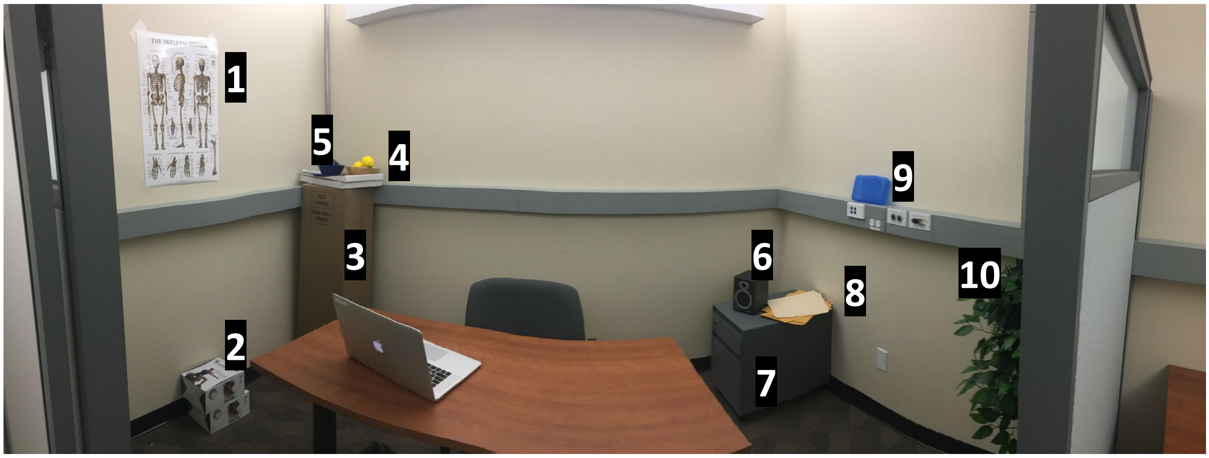
**Acknowledgements** We thank B. Ginn, N. Hall, S. Atwood and M. Nelson for help with data collection; M. Jiang, Y. Mao and B. Chivers for help with data processing; G. Eirich for statistical advice; M. Brucks, K. Duke and A. Galinsky for comments and insights; and J. Pyne and N. Itzikowitz for their partnership.

**Author contributions** M.S.B. supervised data collection by research assistants at the Stanford Behavior Lab in 2016–2021. M.S.B. and J.L. jointly supervised data collection by the corporate partner at the field sites. These data were analysed by M.S.B. and discussed jointly by both of the authors.

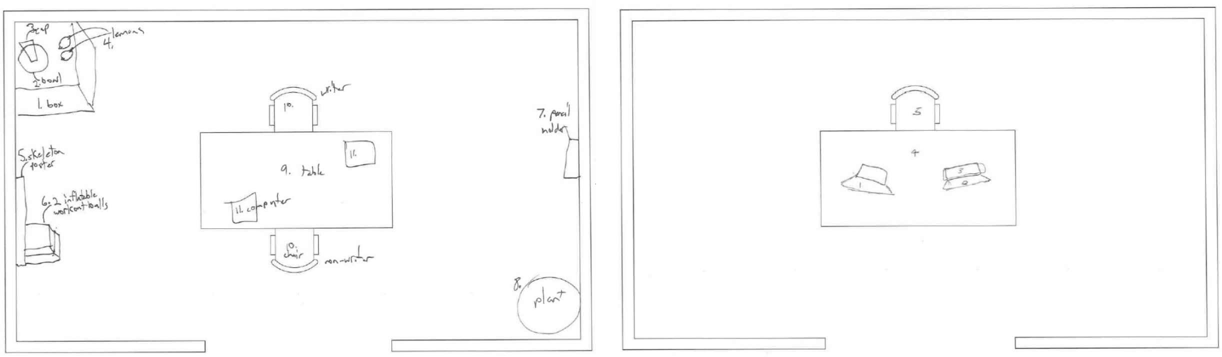
**Competing interests** The authors declare no competing interests.

**Additional information**
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04643-y.
**Correspondence and requests for materials** should be addressed to Melanie S. Brucks.
**Peer review information** *Nature* thanks Brian Uzzi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

a.

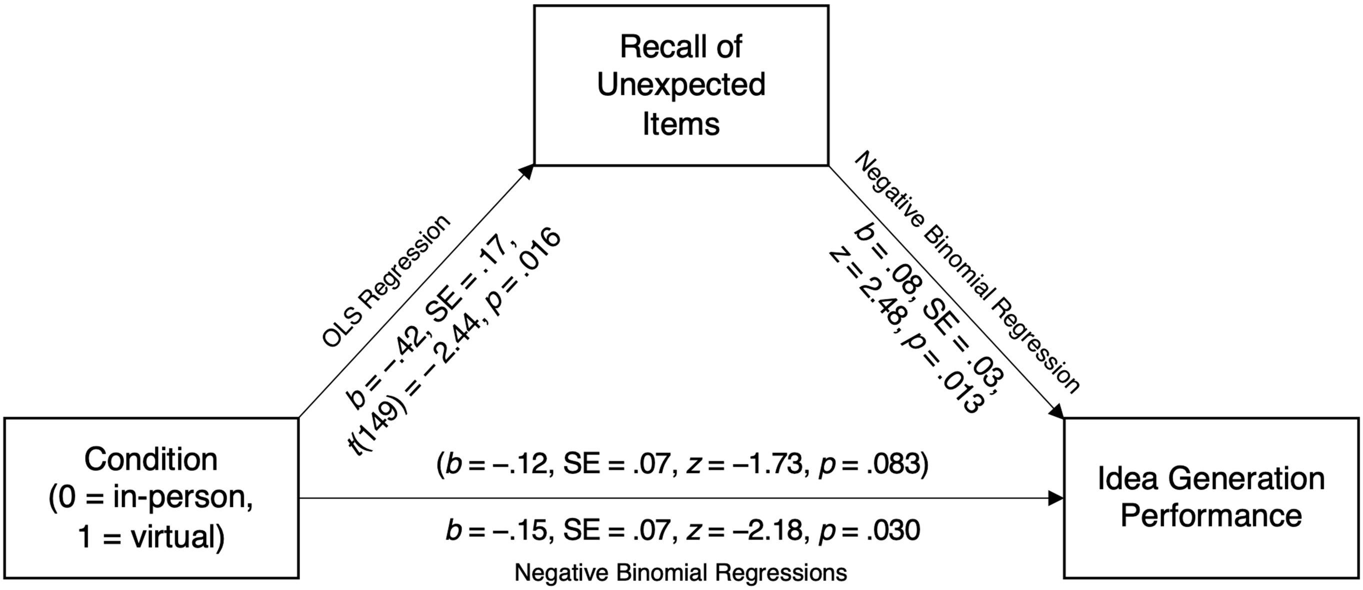



b.

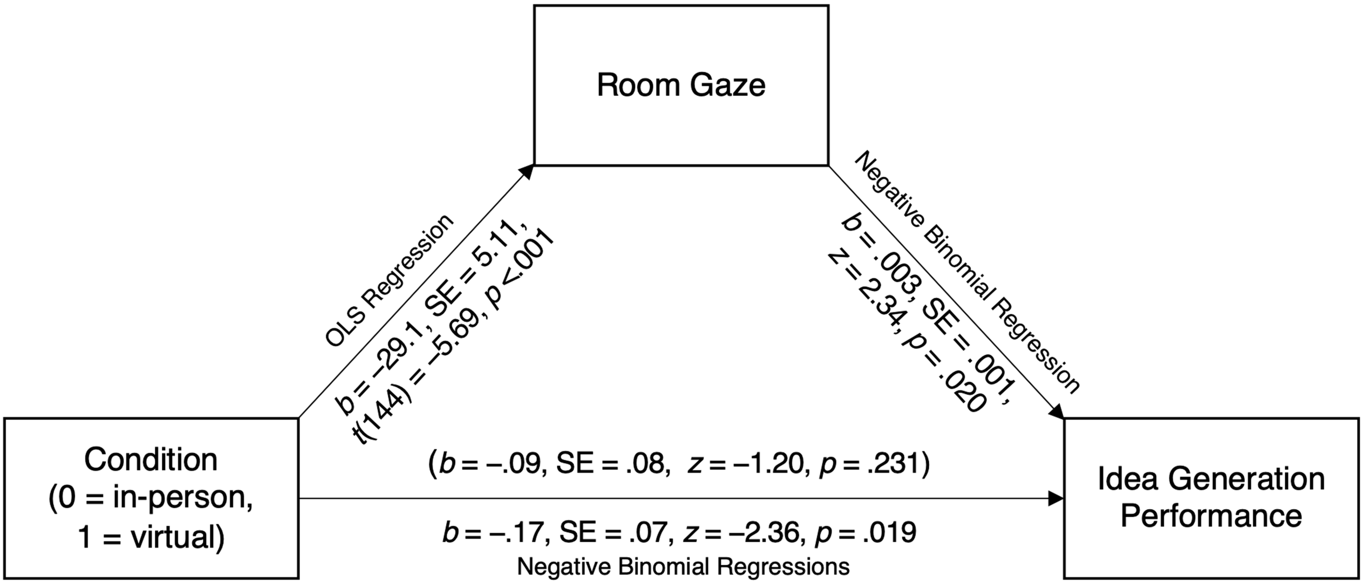



In-person participant example. Virtual participant example.

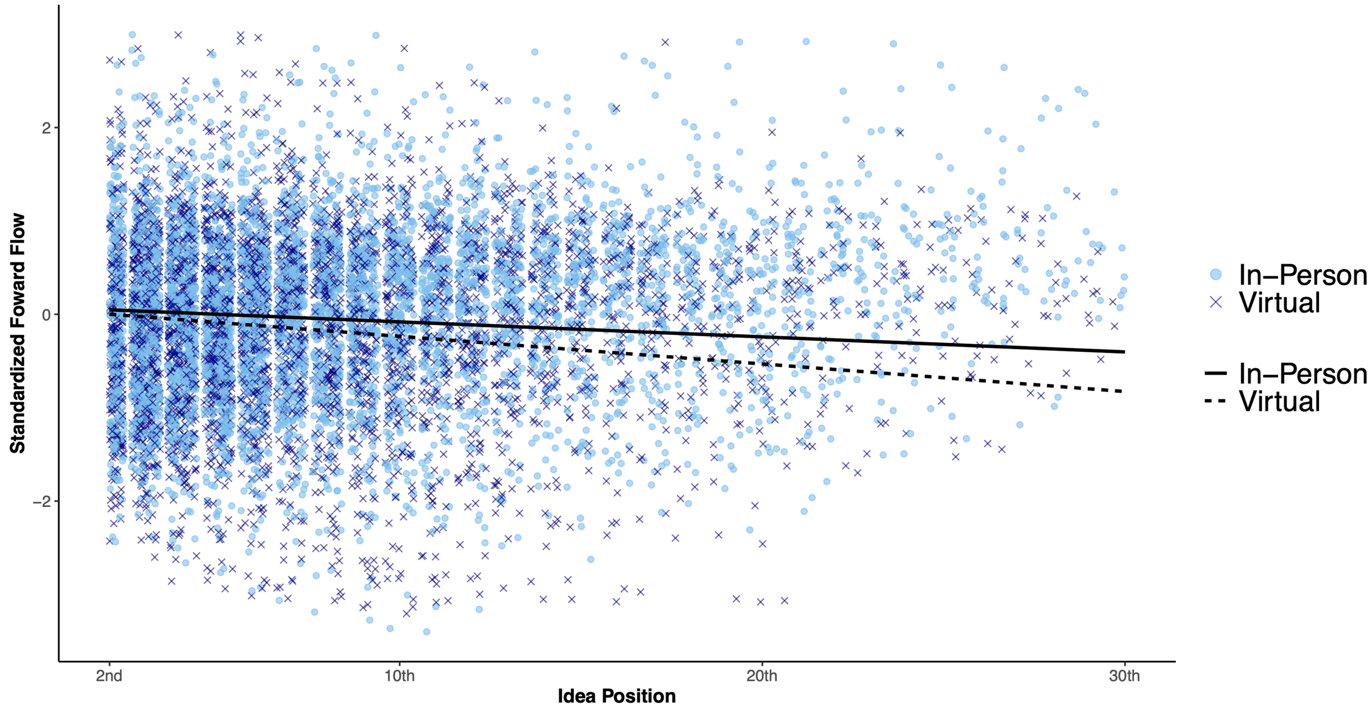
**Extended Data Fig. 1 | Materials and example data for room recall measure in the second batch of data collection in the lab. (a)** Photo demonstrating the prop placement in the lab room. Five props were expected (props consistent with a behavioural lab schema): a filing cabinet, folders, a cardboard box, a speaker, and a pencil box; and five props were unexpected (props inconsistent with a behavioural lab schema): a skeleton poster, a large house plant, a bowl of lemons, blue dishes, and yoga ball boxes. **(b)** Participant example of the data materials. After leaving the lab space, participants recreated the lab room on a piece of paper containing the basic layout of the room and then numbered each element. We then asked participants to list the identity of each element on a Qualtrics survey. A condition- and hypothesis-blind research assistant categorized each listing into one of the ten props and removed any other responses. We then counted how many expected and unexpected props were remembered by each participant.



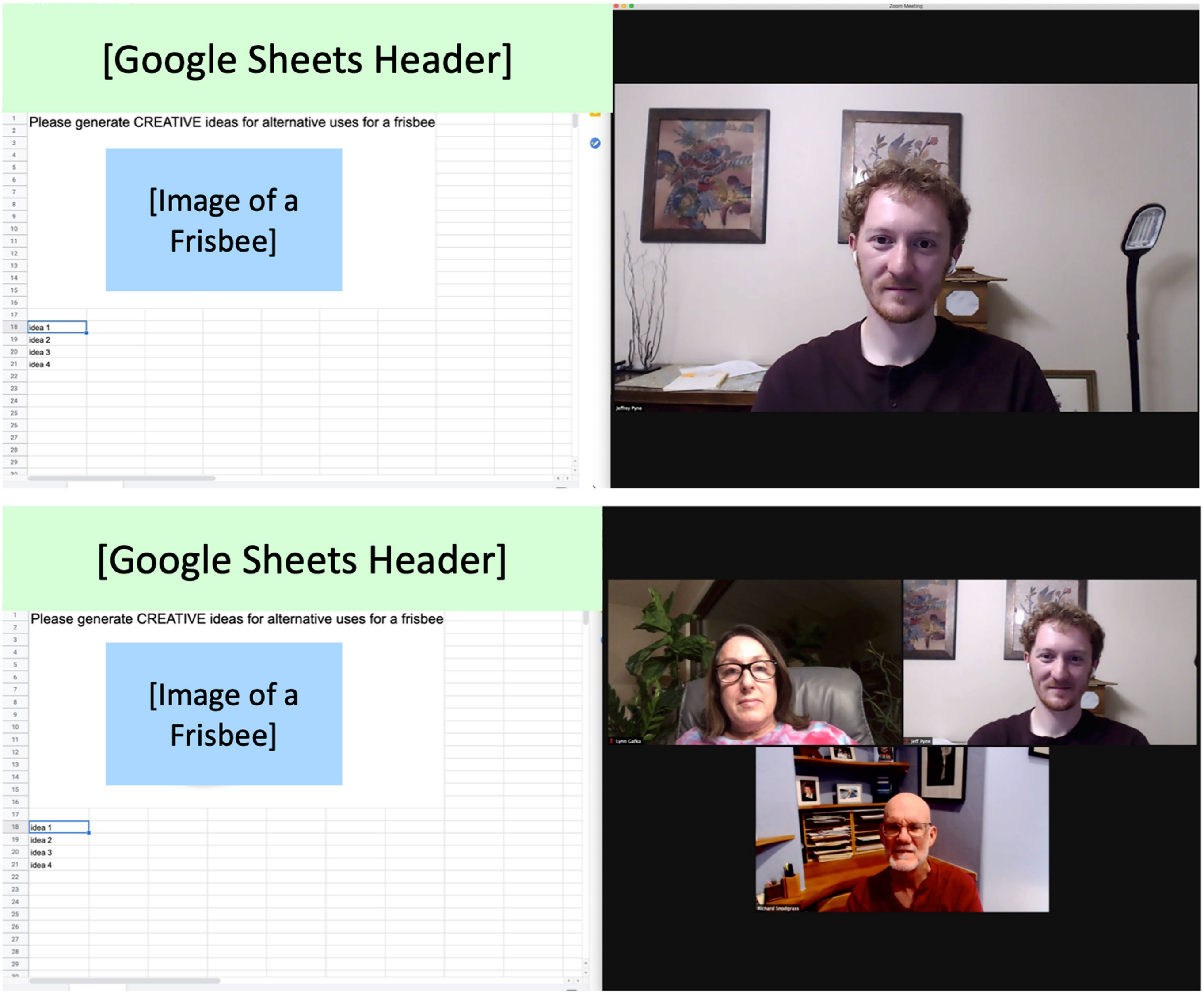

**Extended Data Fig. 2 | Room recall mediates the effect of communication modality on idea generation.** This mediation model demonstrates that virtual participants remembered significantly fewer unexpected props in the experiment room and that this explains the effect of virtual interaction on creative idea generation. We ran an OLS regression for the a-link (communication modality predicting average recall of unexpected items per pair, $n = 151$ pairs, OLS regression, $b = 0.42$, s.e. = 0.17, $t_{149} = 2.44$, $P = 0.016$), and we ran a Negative Binomial regression for the b-link (number of average unexpected items recalled per pair predicting number of creative ideas generated, $n = 151$ pairs, Negative Binomial regression, $b = 0.08$, s.e. = 0.03, $z = 2.48$, $P = 0.013$). A mediation analysis with 10,000 nonparametric bootstraps revealed that recall of the room mediated the effect of modality on creative idea generation (95% confidence intervals of the indirect effect = −0.61 to −0.01). The total effect of modality condition on number of creative ideas generated was significant ($n = 151$ pairs, Negative Binomial regression, $b = 0.15$, s.e. = 0.07, $z = 2.18$, $P = 0.030$), but this effect was attenuated to non-significance when accounting for the unexpected recall mediator ($n = 151$ pairs, Negative Binomial regression, $b = 0.12$, s.e. = 0.07, $z = 1.73$, $P = 0.083$). See Supplementary Information C for model assumption tests of normality and heteroskedasticity. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).




**Extended Data Fig. 3 | Gaze mediates the effect of communication modality on idea generation.** This mediation model demonstrates that virtual participants spent less time looking around the room and that this explains the effect of virtual interaction on creative idea generation. We ran an OLS regression for the a-link (communication modality predicting average room gaze per pair, $n = 146$ pairs, OLS regression, $b = -29.1$, s.e. = 5.1, $t_{144} = 5.69$, $P < 0.001$), and we ran a Negative Binomial regression for the b-link (average room gaze per pair predicting number of creative ideas generated, $n = 146$ pairs, Negative Binomial regression, $b = 0.003$, s.e. = 0.001, $z = 2.34$, $P = 0.020$). A mediation analysis with 10,000 nonparametric bootstraps revealed that recall of the room mediated the effect of modality on creative idea generation (95% confidence intervals of the indirect effect = −1.14 to −0.08). The total effect of modality condition on number of creative ideas generated was significant ($n = 146$ pairs, Negative Binomial regression, $b = 0.17$, s.e. = 0.07, $z = 2.36$, $P = 0.019$), but this effect was attenuated to non-significance when accounting for the room gaze mediator ($n = 146$ pairs, Negative Binomial regression, $b = 0.09$, s.e. = 0.08, $z = 1.20$, $P = 0.231$). See Supplementary Information C for model assumption tests of normality and heteroskedasticity. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).




**Extended Data Fig. 4 | The effect of virtual communication on forward flow across the progression of idea generation.** There was a significant interaction between modality and the position of an idea in the pair's idea sequence on forward flow score across all studies (linear mixed-effect regression, $n = 9966$ idea scores, interaction term: $b = -0.01$, s.e. = 0.01, $t_{358} = -2.09$, $P = 0.038$). At the beginning of the idea generation task, ideas generated by in-person and virtual pairs were similarly connected to past ideas generated by each pair. However, by the eleventh idea, ideas generated by in-person pairs began to exhibit significantly more forward flow (that is, the ideas were less semantically associated) compared to those of virtual pairs (linear mixed-effect regression, $n = 9966$ idea scores, simple effect of modality on forward flow at the $11^{th}$ idea: $b = -0.12$, s.e. = 0.06, $t_{621} = -2.00$, $P = 0.047$). Thus, in-person pairs generate progressively more disconnected ideas relative to virtual pairs. See Supplementary Information D for model assumption tests of normality and heteroskedasticity. We truncated the graph at 30 ideas to provide the most accurate representation of the majority of the data. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).




**Extended Data Fig. 5 | Set-up for group size virtual study.** In the virtual-only study, we randomly assigned participants into groups of 2 or 4 people. Participants worked on a google sheet and were instructed to set up their screen such that half of their screen was the task and the other half of the screen was their zoom window. The self-view was hidden, and participants either saw one partner (2-person condition), or three teammates (4-person condition). Consent was obtained to use these images for publication.

**Extended Data Table 1 | Summary of the analyses in the main text**

| Dependent Variables | Virtual Mean (SD) | In-person Mean (SD) | Comparing the Measure by Condition: Regression Output | Cohen's d |
|---|---|---|---|---|
| **Lab Study: Generation (N = 301 pairs)** | | | | |
| Number of Ideas | 14.74 (6.23) | 16.77 (7.27) | $b = .13$, SE = .05, $z = 2.72$, $p = .007$ | .30 [.07, .53] |
| Number of Creative Ideas | 6.73 (3.27) | 7.92 (3.40) | $b = .16$, SE = .05, $z = 3.14$, $p = .002$ | .36 [.13, .58] |
| Average Creativity | 3.92 (.41) | 3.95 (.39) | $b = .03$, SE = .05, $t(299) = .60$, $p = .547$ | .07 [−.16, .30] |
| **Lab Study: Selection (N = 292 pairs)** | | | | |
| Top Idea | 5.07 (.49) | 5.10 (.46) | $b = .03$, SE = .06, $t(299) = .51$, $p = .608$ | .09 [−.24, .41] |
| Selected Idea | 4.28 (.81) | 4.08 (.84) | $b = .20$, SE = .10, $t(290) = 2.04$, $p = .043$ | −.24 [−.47, .01] |
| Decision Error Score (DES) | .78 (.67) | 1.01 (.77) | $b = .23$, SE = .08, $t(290) = 2.69$, $p = .007$ | .32 [.08, .55] |
| DES Controlling for Number of Ideas | | | $b = .20$, SE = .08, $t(289) = 2.40$, $p = .017$ | |
| **Lab Study: Room Recall (N = 302 participants)** | | | | |
| Total Props Remembered | 2.29 (1.99) | 2.72 (1.76) | $b = .19$, SE = .10, z = 1.97, $p =.049$ | .23 [.00, .46] |
| Unexpected Props Remembered | 1.53 (1.38) | 1.95 (1.38) | $b = .25$, SE = .10, $z = 2.47$ $p = .014$ | .30 [.08, .53] |
| **Lab Study: Eye Gaze (N = 270 participants)** | | | | |
| Partner Gaze | 91.4 (58.3) | 51.7 (52.2) | $b = 39.7$, SE = 6.83, $t(139) = 5.81$, $p < .001$ | .71 [.47, .96] |
| Room Gaze | 32.4 (34.8) | 61.0 (43.1) | $b = 28.8$, SE = 5.10, $t(143) = 5.64$, $p < .001$ | .74 [.49, .99] |
| Task Gaze | 176.1 (64.0) | 186.8 (60.2) | $b = 10.7$, SE = 7.60, $t(268) = 1.40$, $p = .162$ | .17 [−.07, .41] |
| **Field: Generation (N = 745 pairs)** | | | | |
| Number of Ideas | 7.43 (5.17) | 8.58 (6.03) | $b = .14$, SE = .05, $z = 3.13$, $p = .002$ | .21 [.06, .35] |
| Number of Creative Ideas (N = 619) | 3.83 (2.83) | 4.32 (3.18) | $b = .12$, SE = .05, $z = 2.15$, $p = .032$ | .16 [.01, .32] |
| Average Creativity (N = 619) | 2.98 (.40) | 2.99 (.38) | $b = .01$, SE = .03, $t(609) = .30$, $p = .766$ | .02 [−.14, .18] |
| **Field: Selection (N = 619 pairs)** | | | | |
| Top Idea | 3.86 (.56) | 4.01 (.54) | $b = .14$, SE = .04, $t(608) = 3.40$, $p < .001$ | .27 [.11, .43] |
| Selected Idea (N = 591) | 3.05 (.71) | 3.04 (.78) | $b = .004$, SE = .06, $t(582) = .07$, $p = .945$ | .01 [−.16, .17] |
| Decision Error Score (DES, N = 591) | .81 (.76) | .99 (.86) | *Kruskal-Wallis*, $\chi^2(1) = 5.30$, $p = .021$ | .22 [.05, .38] |
| DES Controlling for Number of Ideas | | | $b = .11$, SE = .06, $t(580) = 1.81$, $p = .071$ | |

We used a Negative Binomial regression model for all tests with number of ideas or number of creative ideas as the dependent measure (see Supplementary Information A for model justification and assumption tests). For total props remembered, we used a mixed-effect Negative Binomial Regression and for unexpected props remembered, we used a mixed-effect Poisson regression (Supplementary Information C for test assumptions). For all other dependent measures, when the measure was per participant (e.g., eye gaze), we used a linear mixed-effect regression with random effect of group number. When the measure was per group (e.g., selected idea score), we used an OLS regression model (see Supplementary Information B and C for model assumption tests of normality and heteroskedasticity and for robustness checks). All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).

**Extended Data Table 2 | Summary of idea generation analyses with alternative models**

| Lab | Effect of Modality Condition on Measure | |
|---|---|---|
| Alternative Model | Lab: Number of Ideas (N = 302 pairs) | Lab: Number of Creative Ideas (N = 302 pairs) |
| Quasipoisson | $b$ = .13, SE = .05, $t$(300) = 2.65, $p$ = .008 | $b$ = .16, SE = .05, $t$(300) = 3.16, $p$ = .002 |
| Adjusted Poisson | $b$ = .13, SE = .05, $z$ = 2.67, $p$ = .008 | $b$ = .16, SE = .05, $z$ = 3.17, $p$ = .002 |
| OLS | $b$ = 2.04, SE = .77, $t$(298) = 2.63, $p$ = .009 | $b$ = 1.20, SE = .38, $t$(298) = 3.16, $p$ = .002 |
| Permutation (nonparametric) | Z = 2.58, $p$ = .010 | Z = 3.05, $p$ = .002 |
| **Field** | **Effect of Modality Condition on Measure** | |
| Alternative Model | Field: Number of Ideas (N = 745 pairs) | Field: Number of Creative Ideas (N = 619 pairs) |
| Quasipoisson | $b$ = .14, SE = .05, $t$(744) = 2.79, $p$ = .005 | $b$ = .12, SE = .06, $t$(618) = 2.03, $p$ = .043 |
| Adjusted Poisson | $b$ = .14, SE = .05, $z$ = 2.83, $p$ = .005 | $b$ = .12, SE = .06, $z$ = 2.04, $p$ = .042 |
| OLS | $b$ = 1.15, SE = .41, $t$(737) = 2.79, $p$ = .005 | $b$ = .49, SE = .24, $t$(608) = 2.08, $p$ = .038 |
| Permutation (nonparametric) | Z = 2.80, $p$ = .005 | Z = 2.02, $p$ = .043 |

For the idea generation analysis in the main text, we use a Negative Binomial regression model (see Supplementary Information A for model justification and assumption tests). As a robustness check, we examined the effect of modality on idea generation (both number of ideas and number of creative ideas) using alternative models. We find our results are robust to different regression models. Across all models, there is a significant negative effect of virtual interaction on idea generation. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).

**Extended Data Table 3 | Summary of our examination into the alternative processes by which modality of interaction (in-person vs virtual) could affect the production of creative ideas**

| Alternative Processes Measure | Sample Size | Virtual | In-person | Comparing the Measure by Condition | | Measure Relating to Idea Generation | Effect of Condition when Controlling for Measure |
|---|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Model Output | Cohen's d | | |
| **Subjective Feelings of Connection** | | | | | | | |
| "How much do you like your partner? | 298 participants | 5.69 (1.14) | 5.78 (1.11) | $t(148) = .643, p = .521$ | .09 [–.14, .32] | $z = 1.03, p = .302$ | $z = 2.34, p = .019$ |
| "How similar are you to your partner?" | 298 participants | 3.86 (1.36) | 3.98 (1.33) | $t(147) = .741, p = .460$ | .08 [–.15, .31] | $z = .684, p = .494$ | $z = 2.36, p = .018$ |
| Perceptions of being "in sync" ($\alpha = .82$) | 298 participants | 5.22 (1.09) | 5.40 (.98) | $t(148) = 1.37, p = .173$ | .17 [–.06, .40] | $z = 2.23, p = .026$ | $z = 2.19, p = .029$ |
| Money entrusted to partner | 299 participants | 8.29 (2.65) | 8.42 (2.41) | $t(148) = .40, p = .690$ | .05 [–.17, .28] | $z = 2.73, p = .006$ | $z = 2.12, p = .034$ |
| **Verbal and Nonverbal Behaviors** | | | | | | | |
| *Observer-Rated:* *To what extent did the participant seem…* | | | | | | | |
| Self-conscious | 1378 observations | 3.47 (1.76) | 3.83 (1.77) | $t(129) = 3.12, p = .002$ | .20 [.09, .31] | $z = –.58, p = .559$ | $z = 2.49, p = .013$ |
| Dominant | 1378 observations | 3.45 (1.68) | 3.08 (1.63) | $t(271) = 3.56, p < .001$ | .22 [.11, .33] | $z = –.47, p = .637$ | $z = 2.28, p = .023$ |
| Confident | 1378 observations | 4.47 (1.64) | 4.12 (1.64) | $t(139) = 2.56, p = .012$ | .21 [.11, .32] | $z = .28, p = .783$ | $z = 2.41, p = .016$ |
| Comfortable | 1378 observations | 4.63 (1.64) | 4.37 (1.67) | $t(137) = 2.19, p = .030$ | .16 [.04, .26] | $z = –.26, p = .798$ | $z = 2.31, p = .021$ |
| *Semantic Category Usage (%s)* | | | | | | | |
| Secord Person Singular Pronouns | 516 participants | 4.12 (1.79) | 3.60 (1.49) | $t(256) = 2.55, p = .011$ | .32 [.07, .56] | $z = –1.05, p = .295$ | $z = 2.24, p = .025$ |
| Sad Words | 516 pairs | .16 (.28) | .09 (.20) | $\chi^2 (1) = 5.04, p = .025$ | .26 [.01, .50] | $z = –1.52, p = .128$ | $z = 2.20, p = .028$ |
| **Mimicry** | | | | | | | |
| Facial Action Units: Cosine Similarity | 145 pairs | .86 (.08) | .85 (.09) | $t(143) = .40, p = .688$ | .07 [–.40, .26] | $z = –1.77, p = .076$ | $z = 2.17, p = .030$ |
| Linguistic: Language Style Matching | 258 pairs | .77 (.10) | .77 (.08) | $t(255) = .36, p = .720$ | .04 [–.20, .29] | $z = –2.18, p = .029$ | $z = 2.46, p = .014$ |
| Linguistic: Cosine Similarity | 258 pairs | .91 (.08) | .92 (.07) | $t(256) = .93, p = .354$ | .12 [–.13, .36] | $z = .53, p = .600$ | $z = 2.36, p = .018$ |
| **Communication Coordination** | | | | | | | |
| Self-report | 298 participants | 2.15 (1.47) | 1.83 (1.31) | $t(296) = 2.00, p = .047$ | .23 [.00, .46] | $z = –3.58, p < .001$ | $z = 1.94, p = .052$ |
| Transcript: Crosstalk | 149 pairs | .65 (1.12) | 1.60 (2.40) | $\chi^2(1) = 8.12, p = .004$ | .51 [.19, .84] | $z = –1.01, p = .313$ | $z = 2.47, p = .014$ |
| Transcript: Speaker Switches | 149 pairs | 50.1 (19.3) | 59.6 (21.1) | $t(147) = 2.87, p = .005$ | .47 [.14, .79] | $z = 1.35, p = .176$ | $z = 1.87, p = .062$ |
| **Fear of Evaluation** | | | | | | | |
| How often did you not say an idea because you were worried about what your partner would say about it? | 298 participants | 1.57 (1.12) | 1.52 (1.02) | $t(148) = .42, p = .676$ | .05 [–.18, .28] | $z = –2.50, p = .013$ | $z = 2.35, p = .019$ |
| Perceptions of Criticism ($\alpha = .85$) | 298 participants | 1.48 (.87) | 1.47 (.78) | $t(148) = .11, p = .917$ | .01 [–.21, .24] | $z = –2.81, p = .005$ | $z = 2.44, p = .015$ |
| Perceptions of Warmth ($\alpha = .87$) | 299 participants | 5.89 (.93) | 5.97 (.86) | $t(148) = .72, p = .475$ | .09 [–.14, .31] | $z = .25, p = .800$ | $z = 2.12, p = .034$ |
| Self-Presentation Concerns ($\alpha = .81$) | 299 participants | 2.83 (1.09) | 2.58 (1.04) | $t(148) = 1.90, p = .060$ | .23 [0, .46] | $z = –.06, p = .953$ | $z = 2.14, p = .033$ |
| **Illusion of Productivity** | | | | | | | |
| How much did you enjoy generating ideas? | 599 participants | 5.05 (1.37) | 5.16 (1.34) | $t(367) = .87, p = .387$ | .08 [–.08, .24] | $z = –.28, p = .777$ | $z = 3.13, p = .002$ |
| How satisfied are you with your performance? | 599 participants | 4.84 (1.31) | 4.79 (1.31) | $t(350) = .39, p = .700$ | .04 [–.19, .12] | $z = 1.34, p = .179$ | $z = 3.14$, p = .002 |
| **Dominance** | | | | | | | |
| Verbal Dominance | 258 pairs | .26 (.17) | .21 (.15) | $t(256) = 2.57, p = .011$ | .32 [.07, .57] | $z = –.95, p = .341$ | $z = 2.26, p = .024$ |
| Observer-Rated Dominance | 1,380 observations | 3.45 (1.68) | 3.08 (1.63) | $t(271) = 3.56, p < .001$ | .22 [.11, .33] | $z = –.47, p = .637$ | $z = 2.28, p = .023$ |
| Self-Perceptions of Dominance | 299 participants | 54.7 (17.2) | 56.0 (17.0) | $t(297) = .64, p = .526$ | .07 [–.15, .30] | $z = –.20, p = .844$ | $z = 2.14, p = .032$ |
| Asymmetry in Self-Perceptions of Dominance | 150 pairs | .20 (.20) | .18 (.15) | $t(147) = .43, p = .667$ | .07 [–.39, .25] | $z = –1.34, p = .181$ | $z = 1.99, p = .047$ |
| **Social Loafing** | | | | | | | |
| Did you do a fair share of the work in the idea generation task? | 298 participants | 5.37 (1.28) | 5.49 (1.20) | $t(296) = .836, p = .404$ | .10 [–.13, .32] | $z = .64, p = .521$ | $z = 2.40, p = .017$ |
| Did your partner do a fair share of the work in the idea generation task? | 298 participants | 5.71 (1.21) | 5.90 (1.17) | $t(296) = 1.34, p = .180$ | .16 [.07, .38] | $z = 4.32, p < .001$ | $z = 2.03, p = .042$ |
| **Social Facilitation** | | | | | | | |
| Self-Perceptions of Arousal | 299 participants | 46.2 (20.6) | 41.8 (20.2) | $t(148) = 1.79, p = .076$ | .22 [–.01, .45] | $z = .80, p = .426$ | $z = 2.26, p = .024$ |
| **Production Blocking** | | | | | | | |
| Were you able to voice all of your ideas? | 298 participants | 6.06 (1.23) | 6.32 (1.02) | $t(148) = 1.88, p = .062$ | .23 [.00, .46] | $z = 2.46, p = .014$ | $z = 2.18, p = .029$ |
| **Social Sensitivity** | | | | | | | |
| Ability to "Read" Partner's Personality | 298 participants | .42 (.26) | .41 (.24) | $t(147) = .56, p = .574$ | .07 [–.16, .30] | $z = 1.39, p = .164$ | $z = 2.14, p = .033$ |

Columns two through four examine whether each alternative explanation measure differs by communication modality. For these tests, we ran a linear mixed-effect regression (using the lmerTest package, Bates, Maechler, Bolker, and Walker, 2013) when there was an observation for each participant (e.g., money entrusted to partner). Of note, this package approximates degrees of freedom using Satterthwaite's method. As such, the degrees of freedom vary between these analyses even in cases with the same number of data points. When there was an observation per pair, such as the mimicry measures, we used an OLS regression. For all tests with measures that did not meet the assumptions of a linear model, we instead ran a non-parametric test, which required that we analyse at the group level to account for nested data (Kruskal–Wallis rank-sum test; see Supplementary Information E, F, G, H, I, and J for model assumption tests of normality and heteroskedasticity and for robustness checks). Columns five and six examine the effect of the alternative explanation measure on idea generation. Specifically, column five examines whether this alternative measure relates to the production of creative ideas, and column six examines the effect of modality condition on number of creative ideas when controlling for this alternative measure. For the tests in column six, we used a Negative Binomial regression model (see Supplementary Information A for model justification and assumption tests). All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).

**Extended Data Table 4 | Analyses on Verbal and Nonverbal Behaviours**

| **A. Judges Rating of Non-Verbal Behavior** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **To what extent did the participant engage in the following behaviors…** | α | Virtual Mean (SD) | | In Person Mean (SD) | | *b* | SE | *t* | *p* |
| “act out” something | .93 | 2.52 | (1.77) | 2.58 | (1.77) | 0.05 | 0.14 | 0.37 | 0.710 |
| use body language | .94 | 4.02 | (1.79) | 4.06 | (1.79) | 0.05 | 0.15 | 0.34 | 0.733 |
| have facial expressions | .94 | 4.72 | (1.53) | 4.94 | (1.61) | 0.22 | 0.14 | 1.55 | 0.123 |
| react to their partner | .92 | 4.8 | (1.61) | 4.90 | (1.58) | 0.14 | 0.14 | 1.01 | 0.314 |
| laughing | .97 | 3.49 | (1.97) | 3.54 | (1.96) | 0.07 | 0.22 | 0.33 | 0.742 |
| nodding | .93 | 3.28 | (1.88) | 3.48 | (1.90) | 0.21 | 0.15 | 1.40 | 0.164 |
| touching their face | .99 | 3.05 | (2.18) | 3.07 | (2.26) | 0.08 | 0.25 | 0.34 | 0.733 |
| smiling | .98 | 4.31 | (1.94) | 4.38 | (1.93) | 0.10 | 0.22 | 0.44 | 0.660 |
| **To what extent did the participant seem…** | | | | | | | | | |
| friendly | .95 | 4.94 | (1.57) | 4.04 | (1.57) | 0.15 | 0.15 | 0.99 | 0.323 |
| self-conscious | .85 | 3.36 | (1.74) | 3.63 | (1.74) | 0.19 | 0.11 | 1.74 | 0.084 |
| comfortable | .91 | 4.55 | (1.62) | 4.48 | (1.68) | -0.02 | 0.13 | -0.18 | 0.857 |
| connected to partner | .92 | 4.56 | (1.58) | 4.60 | (1.60) | 0.07 | 0.14 | 0.49 | 0.624 |
| focused | .81 | 4.95 | (1.35) | 4.87 | (1.36) | -0.05 | 0.09 | -0.59 | 0.554 |

| **A. Judges Rating of Verbal Behavior** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **To what extent did the participant seem…** | α | Virtual Mean (SD) | | In Person Mean (SD) | | *b* | SE | *t* | *p* |
| comfortable | .73 | 4.63 | (1.64) | 4.37 | (1.67) | -0.29 | 0.13 | -2.19 | 0.030 |
| friendly | .81 | 5.18 | (1.46) | 5.08 | (1.42) | -0.12 | 0.13 | -0.93 | 0.352 |
| focused | .70 | 5.07 | (1.52) | 4.88 | (1.58) | -0.18 | 0.12 | -1.57 | 0.118 |
| self-conscious | .52 | 3.47 | (1.76) | 3.83 | (1.77) | 0.35 | 0.11 | 3.17 | 0.002 |
| connected with partner | .81 | 4.41 | (1.74) | 4.30 | (1.72) | -0.11 | 0.17 | -0.65 | 0.517 |
| dominant | .64 | 3.45 | (1.68) | 3.06 | (1.63) | -0.4 | 0.11 | -3.56 | <.001 |
| confident | .81 | 4.47 | (1.64) | 4.12 | (1.64) | -0.37 | 0.14 | -2.56 | 0.012 |
| **To what extent did the pair…** | | | | | | | | | |
| Work together as a team | .83 | 4.35 | (1.83) | 4.35 | (1.69) | *Kruskal-Wallis:* $\chi^2(1) = .25$ | | | 0.614 |
| Build on each other's ideas | .81 | 3.95 | (1.89) | 4.07 | (1.89) | 0.1 | 0.19 | 0.51 | 0.613 |
| Intentionally interrupt each other | .37 | 1.93 | (1.43) | 1.99 | (1.40) | 0.06 | 0.08 | 0.69 | 0.494 |
| Accidentally talk over each other | .73 | 2.49 | (1.64) | 2.69 | (1.68) | 0.16 | 0.13 | 1.22 | 0.225 |
| Have moments of silence | .83 | 3.99 | (1.98) | 4.2 | (1.90) | 0.22 | 0.19 | 1.18 | 0.241 |
| **To what extent did the communication seem…** | | | | | | | | | |
| “Chit-chatty” | .66 | 3.12 | (2.08) | 3.12 | (2.01) | -0.04 | 0.17 | -0.24 | 0.810 |
| informal | .46 | 4.73 | (2.13) | 4.74 | (2.08) | -0.02 | 0.14 | -0.13 | 0.893 |
| conversational | .74 | 4.36 | (2.33) | 4.27 | (2.29) | -0.15 | 0.21 | -0.72 | 0.470 |
| focused | .68 | 5.33 | (2.07) | 5.16 | (2.02) | -0.18 | 0.15 | -1.15 | 0.251 |
| **To what extent did the participant engage in the following behaviors…** | | | | | | | | | |
| vocal pitch fluctuation | .63 | 3.41 | (1.52) | 3.45 | (1.61) | 0.01 | 0.11 | 0.11 | 0.912 |
| react to their partner | .85 | 4.83 | (1.75) | 4.8 | (1.69) | -0.05 | 0.17 | -0.28 | 0.778 |
| volume fluctuation | .63 | 3.68 | (1.62) | 3.69 | (1.67) | 0 | 0.12 | -0.03 | 0.974 |

(a) Summary of the effect of communication modality on nonverbal behaviour (objective judges' scores of muted videos). (b) Summary of the effect of communication modality on verbal behaviour (objective judges' scores of videos with sound). For these regressions, we ran linear mixed-effect regression (using the lmerTest package, Bates, Maechler, Bolker, and Walker, 2013) and included a random effect of judge and group number because each judge evaluated five videos and we had two videos per pair (one for each participant). See Supplementary Information F for model assumption tests of normality and heteroskedasticity and for robustness checks. When the measures did not meet the assumptions of a linear model, we instead ran a non-parametric test (Kruskal–Wallis rank-sum test). For non-verbal behaviour, $n$=1,676 observations; for verbal behaviour, $n$=1,380 observations. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).

**Extended Data Table 5 | Summary of group size virtual study results**

| Task Performance | 2-person Mean (SD) | 4-person Mean (SD) | Cohen's d | Regression Statistics | | | |
|---|---|---|---|---|---|---|---|
| | | | | b | SE | test statistic | p-value |
| **Idea Generation** | | | | | | | |
| Number of ideas per person | 9.50 (3.20) | 5.56 (1.75) | −1.55 [−2.09, −1.01] | *Kruskal-Wallis:* $\chi^2(1) = 35.2$ | | | $p < .001$ |
| Number of creative ideas per person | 5.38 (2.21) | 3.13 (1.11) | −1.31 [−1.81, −1.79] | *Kruskal-Wallis:* $\chi^2(1) = 26.2$ | | | $p < .001$ |
| Number of ideas (with nominal groups of 4) | 30.67 (8.20) | 22.23 (6.98) | −1.15 [−1.77, −.51] | −.16 | .03 | $z = -5.21$ | $p < .001$ |
| Number of creative ideas (with nominal groups of 4) | 18.10 (5.83) | 12.51 (4.44) | −1.14 [−1.75, −.50] | −.18 | .03 | $z = -5.11$ | $p < .001$ |
| Average creativity of all ideas | 3.66 (.21) | 3.68 (.22) | .08 [−.35, .50] | .02 | .05 | $t(83) = .34$ | $p = .731$ |
| Standard deviation of creativity of ideas | .80 (.11) | .77 (.12) | −.27 [−.69, .16] | −.03 | .02 | $t(83) = -1.21$ | $p = .228$ |
| **Idea Selection** | | | | | | | |
| Decision Error Score | 1.29 (.88) | 1.05 (.79) | −.29 [−.72, .14] | −.24 | .18 | $t(83) = -1.32$ | $p = .190$ |
| Creativity Score of Selected Idea | 3.88 (.96) | 4.10 (.79) | .25 [−.18, .68] | .22 | .19 | $t(83) = 1.14$ | $p = .256$ |
| Creativity Score of Top Idea | 5.17 (.57) | 5.15 (.46) | −.04 [−.47, .38] | −.02 | .11 | $t(83) = -.19$ | $p = .849$ |

| Process Measure | 2-person Mean (SD) | 4-person Mean (SD) | Cohen's d | Regression Statistics |
|---|---|---|---|---|
| **Evaluation Apprehension** ($\alpha = .84$)<br>1. How often did your group member/s criticize your ideas?<br>2. How often did you not say an idea because you were worried about what your group member/s would think of it?<br>3. Were you reluctant in offering an idea for fear of criticism from your group member/s?<br>4. Did fear of possible disapproval from your group member/s make you withhold any ideas?<br>5. How often did you sense a certain disapproval from your group member/s, even if no overt criticism was expressed? | 1.51 (.78) | 1.67 (.97) | .18 [−.08, .45] | *Kruskal-Wallis:* $\chi^2(1) = 6.17$, $p = .013$ |
| **Satisfaction with Performance** ($\alpha = .88$)<br>1. How satisfied are you with your group's idea generation performance?<br>2. Do you think your group performed well during idea generation? | 5.16 (1.43) | 5.49 (1.14) | .27 [.00, .53] | $b = .32$, SE = .19, $t(90) = 1.77$, $p = .080$ |
| **Social Loafing** ($\alpha = .74$)<br>1. How much did the success of your group depend upon you, personally?<br>2. To what extent did you feel high degree of personal ownership over the ideas generated in this task?<br>3. To what extent do you feel the ideas that were generated during the task were your ideas?<br>4. How much effort did you put in to coming up with ideas for this task? | 4.61 (.73) | 4.25 (1.07) | −.38 [−.65, −.11] | *Kruskal-Wallis:* $\chi^2(1) = 10.28$, $p = .001$ |
| **Production Blocking** ($\alpha = .80$)<br>1. Were you able to voice all of your ideas?<br>2. Did you say every idea that occurred to you during the group brainstorming session? | 5.95 (1.32) | 5.72 (1.50) | −.16 [−.43, .10] | $b = -.24$, SE = .19, $t(252) = -1.22$, $p = .223$ |
| **Conversation Coordination** ($\alpha = .72$)<br>1. How often did your group struggle with determining who would speak when?<br>2. How difficult was it to determine who would speak next during idea generation?<br>3. How often did two people speak at once accidentally?<br>4. How often did you feel like you were talking over each other?<br>5. How often did you feel like there were awkward silences during idea generation? | 2.02 (.87) | 2.30 (.88) | .32 [.06, .59] | $b = .28$, SE = .13, $t(100) = 2.26$, $p = .026$ |

We used an OLS regression model for all tests with number of ideas per person and number of creative ideas per person as the dependent measure as these measures are not discrete integers. We used a Negative Binomial regression model for number of creative ideas and number ideas for nominal groups of 4 (where we randomly assigned two 2-person groups into nominal groups of 4 and removed redundant ideas). We used an OLS regression model for all other tests of task performance. We used a linear mixed-effect regression model for all other participant-level dependent measures, including a random effect for group number. See Supplementary Information M for model assumption tests of normality and heteroskedasticity and for robustness checks. When the measures did not meet the assumptions of a linear model, we instead ran a non-parametric test (Kruskal–Wallis rank-sum test). In these analyses, $n$=85 groups and $n$=254 participants. All tests are two-tailed and there were no adjustments made for multiple comparisons (for a discussion of our rationale, see Supplementary Information S).

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<br>*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<br>*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 overdisperson 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.