nature portfolio

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

Supplementary information

Virtual communication curbs creative idea generation

In the format provided by the
authors and unedited

SUPPLEMENTARY INFORMATION

Title: Virtual Communication Curbs Creative Idea Generation

Authors: Melanie S. Brucks[1*], Jonathan Levav[2]

Affiliations:

[1] Marketing division, Columbia Business School
Columbia University in the City of New York
665 West 130th Street, New York, NY 10027, USA

[2] Marketing division, Stanford Graduate School of Business
Stanford University
655 Knight Way, Stanford, CA 94305, USA

Additional Information

Correspondence and requests for materials should be addressed to Melanie Brucks (mb4598@columbia.edu).

TABLE OF CONTENTS

A. Testing Model Assumptions for Idea Generation Analyses

For the main analyses in the paper, the dependent variables are number of ideas generated and number of creative ideas generated by each pair. These are count measures consisting of non-negative integers. The Poisson family of regression models are typically used to model count measures[1].

There are multiple models within the Poisson family, but the two most common models in psychology literature use the standard Poisson distribution or the Negative Binomial distribution[2]. The Poisson regression model assumes that the mean and variance are equal. In most real-world data, this assumption is not met, and the variance is greater than the mean. This is termed "overdispersion." The Negative Binomial regression model can account for overdispersion and is thus more appropriate when count data are overdispersed[3].

To identify the most appropriate regression model for our data, we first examined if our data are overdispersed. As with many real-world datasets, we find that our data are overdispersed (overdispersion test in the AER R package; Lab study: *total ideas*: $\alpha = 1.80$, $z = 5.46$, $p < .001$; *creative ideas*: $\alpha = .47$, $z = 3.34$, $p < .001$, field study: *total ideas:* $\alpha = 2.88$, $z = 6.98$, $p < .001$, *creative ideas:* $\alpha = 1.17$, $z = 5.63$, $p < .001$)[1] and, as a result, do not meet the restrictive assumption of the Poisson distribution. Thus, a Negative Binomial distribution model is theoretically more appropriate. Supporting this conclusion, the Akaike information criterion (AIC) was larger for the Poisson regression model (Lab study: *total ideas*: 2,164; *creative ideas*: 1,563; field study: *total ideas:* 5,179, *creative ideas:* 3,055) than the Negative Binomial regression model (Lab study: *total ideas*: 1,954; *creative ideas*: 1,542; field study: *total ideas:* 4,347, *creative ideas*: 2,888) for all four tests. Further, Vuong's (1989) test finds that that the Negative Binomial regression model is a significantly better fit than the Poisson regression model for all four tests (Lab study: *total ideas*: $z = 4.19$, $p <.001$, *creative ideas*: $z = 1.95$, $p =.026$, field study: *total ideas:* $z = 6.46$, $p <.001$, *creative ideas:* $z = 4.25$, $p <.001$)[4]. Thus, we report results from Negative Binomial regression models in the main text and supplement whenever the number of ideas or number of creative ideas is the dependent variable.

In addition, we also conducted the analyses using a variety of other regression models as a robustness check. Specifically, we examined the effect of modality on number of ideas and number of creative ideas generated using the Quasipoisson regression model, the adjusted Poisson regression model, OLS regression model, and a non-parametric permutation test. Across all models for all tests, we observe a significant effect of communication modality on idea generation. See Extended Data Table 2 for results.

B. Testing Model Assumptions for Idea Selection Models

To ensure that our selection data meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6].

Our lab data do not violate the homogeneity of variance assumption: the variances between modality conditions were not significantly different from each other for either selection measure (*error score*: $F(1, 290) = 3.51$, $p = .062$; *creativity of the selected idea*: F(1, 290) = .012, $p = .913$). However, the distributions of both measures were significantly non-normal (Shapiro-Wilk normality test: *error score*: $W = .93$, $p <.001$; *creativity of the selected idea*: $W = .98$, $p < .001$). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the lab analyses using a non-parametric test (which makes no assumptions regarding the underlying distribution). These non-parametric tests support similar conclusions as the linear regressions reported in the paper (Kruskal-Wallis rank sum test, N = 292 pairs, *difference score*: χ^2 (1) = 6.10, $p = .014$; *creativity of selected idea*: Kruskal-Wallis rank sum test, N = 292 pairs, χ^2 (1) = 3.76, $p = .052$).

In our field data, the variances between modality conditions were significantly different for the error score ($F(1, 589) = 5.29$, $p = .022$). Thus, we report results from a Kruskal-Wallis rank sum test. However, the results do not differ meaningfully from a linear mixed regression (the effect of condition on error score, linear mixed effect regression, N = 591 pairs, $b = .17$, SE = .07, $t(582) = 2.57$, $p = .010$). The variances were not significantly different between modality conditions for the selected idea score (*F*(1, 589) = 2.72, *p* = .100), thus a linear regression is appropriate. Of note, we did not check for normality because our sample size (N = 568 pairs) greatly surpasses the 100 sample-size threshold.

C. Testing Model Assumptions for Process Analyses

To ensure that our process data meet the assumptions of a linear regression, we evaluated whether the measures' variances were homogenous by conducting the Levene's test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6].

Eye Gaze. Our measure of time spent looking at the task did not significantly violate the homogeneity of variance assumption (Levene's Test: $F(1, 144) = .02$, $p = .875$) and was not significantly non-normal (Shapiro-Wilk normality test: $W = .98$, $p = .10$). Thus, we only reported the results from a linear regression model in the main text.

Our measure of time spent looking at one's partner did not significantly violate the homogeneity of variance assumption (Levene's Test: $F(1, 144) = .07$, $p = .796$) but was significantly non-normal (Shapiro-Wilk normality test: $W = .95$, $p < .001$). The normality result should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the analysis of condition predicting time spent looking at one's partner using a non-parametric test (which makes no assumptions regarding the underlying distribution). The results of this test are similar to the results reported in the main text (Kruskal-Wallis rank sum test, N = 146 pairs, $\chi^2 (1) = 32.28$, $p < .001$).

Our measure of time spent looking at the room significantly violated the homogeneity of variance assumption (Levene's Test: $F(1, 144) = 6.13$, $p = .014$) and was significantly non-normal (Shapiro-Wilk normality test: $W = .92$, $p < .001$). Importantly, the results examining the effect of condition on time spent looking at the room using a non-parametric test were not meaningfully different from the results of the linear regression model (Kruskal-Wallis rank sum test, N = 146 pairs, χ^2 (1) = 29.75, p < .001).

Memory. Our measures of number of unexpected and total props remembered are count measures consisting of non-negative integers. The Poisson family of regression models are typically used to model count measures[1]. To identify the most appropriate regression models for our data, we first examined if our data are overdispersed. We find that the memory for total props is significantly overdispersed (overdispersion test in the AER R package, α = .43, z = 3.37, p <.001), but the memory for unexpected props is not significantly overdispersed (α = .11, z = 1.45, p =.074). Thus, a Negative Binomial distribution model is theoretically more appropriate for total prop memory, and a Poisson distribution is more appropriate for the unexpected prop recall.

Supporting this conclusion, the Akaike information criterion (AIC) was larger for the Poisson regression model compared to the Negative Binomial regression model for total prop recall (NB: 1,189, P: 1,206), and the AIC was (only slightly) larger for the Negative Binomial regression model compared to the Poisson regression model for unexpected prop recall (NB: 1,009, P: 1,008). Further, Vuong's (1989) test finds that that the Negative Binomial regression model is a significantly better fit than the Poisson regression model for all total prop recall (z = 1.90, p = .028), but that the Negative Binomial regression model is not a significantly better fit than the Poisson regression model for all unexpected prop recall (z = .64, p = .261)[4]. Thus, we report results from Negative Binomial regression model for total prop recall and Poisson regression model for unexpected prop recall.

D. Latent Semantic Analysis

We used latent semantic analysis (LSA) to calculate how narrow and closely connected each pair's idea stream was[9]. LSA leverages co-occurrence in a text to quantify similarity. Words that co-occur often, such as "plastic" and "water," have smaller semantic distance than words that do not co-occur often, such as "plastic" and "medieval." Forward flow quantifies how ideas unfold over time in sequence by using LSA to compute the semantic distance of each idea from all the preceding ideas in the "thought stream" of an individual or group[17]. A high forward flow score indicates that an idea departs significantly from the prior ideas. A low forward flow score indicates that an idea is semantically similar to prior ideas. For example, consider the following two idea sequences for uses for a Frisbee: (a) plate, food cover, cup, serving platter; (b) serving platter, armor, dog sled, hat. The first sequence has low forward flow because each idea is semantically similar (food-related in this example), whereas the second sequence has higher forward flow because each idea semantically departs from prior ideas.

To calculate the "forward flow" of each idea in our study, we wrote custom Python (version 3.8.1) code. The "forward flow" calculation begins with a seed word, and the forward flow of the first idea is determined by calculating the semantic distance between the first idea and this seed word. The seed word for the first batch of data collection in the lab was "Frisbee," and the seed word term for the second batch of data collection in the lab was "bubble wrap." Because there was no specific topical context in the field study (i.e., any product ideas were permitted), we used the first idea as the seed word. To do this, we removed the score of the first idea from the analysis. In the lab study, when calculating forward flow, we used the same corpus as the original corpus used to validate the "forward flow" measure[9]. For the field study, due to the highly technical nature of the ideas, we used a Wikipedia corpus, which is more comprehensive[10]. When no words in an idea were identified in the corpus, we treated the score for that idea as a missing value and did not include it in the calculation for subsequent ideas.

When thinking is more focused, the idea sequence should "stay on topic," which would result a lower forward flow score (i.e., ideas being more semantically related to past ideas). Alternatively, when thinking is less focused, the idea sequence should jump around from topic to topic, which would result in higher forward flow (i.e., ideas being less semantically related to past ideas). We hypothesized that, as participants communicate virtually and become increasingly focused, forward flow should decrease relative to in-person pairs. To test this, we collapsed across study and ran a mixed effect linear regression, standardized forward flow score as the dependent variable, and communication modality, the sequence position of the idea, and their interaction as predictors, study and total number of ideas generated as a covariate, and pair as a random intercept and with a random slope of sequence position. There was a significant communication modality × idea sequence position interaction (linear mixed effect regression, N = 9,966 ideas, *b* = .014, SE = .06, *t*(358) = 2.08, *p* = .038). Spotlight analyses demonstrate that the difference in forward flow between modality conditions was not significant at the second idea (selected because there was no first idea score in the field studies, simple slope of linear mixed effect regression, N = 9966, *b* = .002, SE = .05, *t*(934) = .03, *p* = .973). However, in-person pairs exhibited significantly higher flow starting at the 11th idea (simple slope of linear mixed effect regression, N = 9966, *b* = .12, SE = .06, *t*(621) = 2.00, *p* = .047).

Of note, we tested the assumptions of a linear regression model to ensure it was the appropriate model. Our measure of forward flow did not violate the homogeneity of variance assumption (Levene's Test: *F*(1, 1039) = 1.63, *p* = .202). Further, we did not confirm normality because the central limit theorem states that, at sufficiently large sample sizes (>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8]. Given that we had approximately 10,000 observations and 1,041 pairs, our sample greatly surpasses the 100 sample-size threshold and thus it is unnecessary to check for normality.

E. Trust and Social Connection Analysis

This section covers the methods and analyses used to examine the effect of communication modality on social connection. We first examined the effect of condition on self-report measures collected via surveys at the end of the lab study. Specifically, in the first batch of data collection, we asked each participant to how in sync they felt with their partner (“Did you feel "in sync" with your partner while generating ideas?” “Did you feel like you and your partner were "on the same page" during the task?", and “Did you and your partner work well together?”, from 1 = *not at all* to 7 = *very much*), how much the liked their partner (from 1 = *not at all* to 7 = *very much*), and how similar they were to their partner (from 1 = *not at all* to 7 = *very much*).

Before running analyses, to ensure that our data meet the assumptions of a linear regression, we evaluated whether the data’s variance was homogenous by conducting the Levene’s test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. We do not significantly violate the homogeneity of variance assumption: the variances of each measure did not significantly differ by condition (*in-sync*: $F(1, 296) = .09$, $p = .761$; *like*: $F(1, 296) = .21$, $p = .647$; *similarity*: $F(1, 296) = .06$, $p = .799$). However, all three measures are significantly non-normal (Shapiro-Wilk normality test for *in-sync*: $W = .96$, $p < .001$; *like*: $W = .87$, $p <.001$; and *similarity*: $W = .94$, $p <.001$). These normality results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the analyses using a non-parametric test (which makes no assumptions regarding the underlying distribution) and include those results with the linear regression model results below.

To examine the effect of modality on each of these measures, we ran linear mixed effect regressions with the measure as the dependent variable, condition as the independent variable, and group number as a random effect. Virtual pairs did not significantly differ from in-person pairs in their perceptions of how in sync they felt ($M_{virtual}$ = 5.22, SD = 1.09, $M_{in\text{-}person}$= 5.40, SD = .98, linear mixed effect regression, N = 298 participants, *b* = .18, SE = .13, *t*(148) = 1.37, *p* = .173, Cohen's *d* = .17, 95% CI: [–.06, .40]), their liking of their partner ($M_{virtual}$ = 5.69, SD = 1.14, $M_{in\text{-}person}$= 5.78, SD = 1.11, linear mixed effect regression, N = 298 participants, *b* = .09, SE = .14, *t*(148) = .64, *p* = .521, Cohen's *d* = .08, 95% CI: [–.15, .31]) or perceived similarity ($M_{virtual}$ = 3.86, SD = 1.36, $M_{in\text{-}person}$= 3.98, SD = 1.33, linear mixed effect regression, N = 298 participants, *b* = .12, SE = .16, *t*(147) = .741, *p* = .460, Cohen's *d* = .09, 95% CI: [–.14, .32]). Lastly, we examined if the significance of the effect of communication modality on idea generation altered when accounting for these measures. We find that controlling for these measures does not attenuate the negative effect of condition on idea generation (negative binomial regression, N = 150 pairs, *b* = .16, SE = .07, *z* = 2.18, *p* = .029). Non-parametric tests (which do not assume a type of distribution) reached the same statistical conclusions (Kruskal-Wallis rank sum test, N = 150 pairs, *in-sync*: χ^2 (1) = 1.59, *p* = .207; *like*: χ^2 (1) = .91, *p* = .341; *similar*: χ^2 (1) = .12, *p* = .725).

We next examined social connection via in an economic trust game at the end of the second batch of data collection in the lab. In this trust game, each participant individually decided whether to trust their partner with a $10 bonus (if they entrusted all $10 to their partner, the money would triple to $30, but the partner would then choose how much to keep and how much to give back to the other partner, see Lab Experiment: Stimulus 2 in Methods for procedural details). We ensured that the variance in the measure of money entrusted did not significantly differ by modality (Levene's Test, $F(1,297) = .22$, $p = .641$). However, the measure of money entrusted is significantly non-normal (Shapiro-Wilk normality test, $W = .67$, $p <.001$). Although this is not of great concern (as mentioned above), as a robustness check, we will also report the results of a non-parametric test below.

To examine if trust differed by communication modality, we ran a linear mixed effect regression with money entrusted as the dependent variable, condition as the independent variable, and group number as a random effect. We found that the amount participants entrust to their partner did not significantly differ by communication modality ($M_{virtual} = 8.29$, SD = 2.65, $M_{in\text{-}person}= 8.42$, SD = 2.41, linear mixed effect regression, N = 299 participants, $b = .13$, SE = .33, $t(148) = .40$, $p = .690$, Cohen's $d = .05$, 95% CI: [–.17, .28]). A non-parametric test replicates this result (Kruskal-Wallis rank sum test, N = 150 pairs, χ^2 (1) = .15, $p = .703$).

Next, we examined if trust relates to ideation by running a negative binomial regression with number of creative ideas as the dependent variable and average money entrusted as the independent variable. We found that amount of money entrusted to one's partner significantly and positively correlates with the number of creative ideas the pair generated, replicating prior work that suggests that trust facilitates ideation[11,12] (negative binomial regression, N = 150 pairs, $b = .05$, SE = .02, $z = 2.73$, $p = .006$). Lastly, and most importantly, we found that controlling for money entrusted did not significantly attenuate the effect of modality on idea generation (negative binomial regression, N = 150 pairs, $b = .15$, SE = .07, $z = 2.12$, $p = .034$).

These results provide initial evidence that differences in trust and connection do not underly the effect we observe.

F. Video Analysis

Method

In the second batch of data collection in the lab, we video-recorded participants' interactions. To evaluate if there are any differences in social behaviors by condition, we recruited online judges on Prolific Academic to watch three randomly assigned clips of a participant generating ideas with their partner. Each judge scored each clip on 13 nonverbal dimensions (including concrete behaviors such as body language and smiling and more abstract behaviors such as how comfortable the participant seemed, see Extended Data Table 4a) or 19 verbal dimensions (including concrete behaviors such as laughing and volume fluctuation and more abstract behaviors such as "working together as a team," see Extended Data Table 4b).

Specifically, we trimmed the video of each participant interacting with their partner to a 30-second segment (starting 30 seconds in, i.e., from :30 to 1:00) because prior research reveals that 30 seconds is a sufficient amount of time to develop systematic and lasting impressions about another person and their behavior[13]. We created two sets of videos: one set was muted, to focus specifically on nonverbal behavior, and the other set included sound.

We recruited 658 judges (303 men, 337 women, 18 nonbinary; M_{age} = 31.7; SD_{age} = 12.15) to evaluate the muted videos and 546 judges (259 men, 280 women, 7 nonbinary; M_{age} = 28.5; SD_{age} = 10.21) to evaluate the videos with sound. Out of 302 participants, 281 videos of participants videos were scored. This is because nine videos were not saved, six videos cut off participants' eyes, four videos were too dark to reliably code, and two videos were corrupted and could not load. We sought to recruit five judges per video, with each judge evaluating three randomly assigned videos. Due to attrition, a small number of videos were evaluated four rather than five times (out of 281 videos, 65 videos with sound and 7 muted videos) and a few less than four times (7 videos with sound).

Before viewing the videos, judges were informed that we had run a study on group interaction and that they would be observing a video recording taken from the task computer of one of the participants in a pair. Judges of the muted videos were informed that the videos were muted, that we were interested in the *nonverbal* behavior the participant engages in and that they should pay attention to the participant's facial and body expressions. Judges of the videos with sound were informed that participants were generating uses for bubble wrap, that we were interested in the *verbal* behavior the participant engages in and that they should pay attention to what the participant says and how they say it. Importantly, judges were not informed about the experimental condition or hypotheses and were required to watch the 30-second video in its entirety before scoring the participant on any dimensions. At the end of each video, judges indicated if they couldn't hear or see the face of the participant for at least half the video, and we excluded any video where at least three of the five judges indicated "yes" on these items (6 muted videos, 7 videos with sound).

Results

Extended Data Table 4a summarizes the results of the nonverbal behaviors and Extended Data Table 4b summarizes the results of the verbal behaviors scored by judges. Of note, the interrater reliabilities across dimensions are quite high, and with the exception of a couple measures (how informal the participant seemed and how often the participants intentionally interrupted each other), they meet the Cicchetti and Sparrow criteria for satisfactory agreeability[14]. This suggests that there is an interpretable signal in the videos.

Before running analyses, to ensure that our data meet the assumptions of a linear regression, we tested the homogeneity of variance using the Levene's test. The Levene's test revealed that for all categories but one (working together: $F(1, 1378) = 6.14$, $p = .013$), variance was not significantly different by condition. For the "working together" category, we ran a non-parametric test. Of note, we did not confirm normality for each measure because the central limit theorem states that, at sufficiently large sample sizes (>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8]. Given that we had multiple judges (and over 1,000 observations), our sample greatly surpasses the 100 sample-size threshold and thus it is unnecessary to check for normality.

To examine whether communication modality affects the verbal and non-verbal behaviors of participants, we ran linear mixed effect regressions with the measure as the dependent variable, condition as the independent variable, and group number and observer as random effects. Across the analyses, the only behaviors that were significantly different were how self-conscious ($M_{virtual}$ = 3.47, SD = 1.76, $M_{in\text{-}person}$= 3.83, SD = 1.77, linear mixed effect regression, N = 1380 ratings; 274 participants, $b = .35$, SE = .11, $t(129) = 3.12$, $p = .002$, Cohen's $d = .20$, 95% CI: [.09, .31]), comfortable ($M_{virtual}$ = 4.63, SD = 1.64, $M_{in\text{-}person}$= 4.37, SD = 1.67, linear mixed effect regression, N = 1380 ratings; 274 participants, $b = .29$, SE = .13, $t(139) = 2.19$, $p = .030$, Cohen's $d = .16$, 95% CI: [.05, .26]), dominant ($M_{virtual}$ = 3.45, SD = 1.68, $M_{in\text{-}person}$= 3.08, SD = 1.63, linear mixed effect regression, N = 1380 ratings; 274 participants, $b = .40$, SE = .11, $t(271) = 3.56$, $p < .001$, Cohen's $d = .22$, 95% CI: [.11, .33]), and confident ($M_{virtual}$ = 4.47, SD = 1.64, $M_{in\text{-}person}$= 4.12, SD = 1.64, linear mixed effect regression, N = 1380 ratings; 274 participants, $b = .37$, SE = .14, $t(139) = 2.56$, $p = .012$, Cohen's $d = .21$, 95% CI: [.11, .32]) the participant seemed.

We next examined whether these measures were associated with idea generation performance of each pair. We found that none of these measures significantly correlated with number of creative ideas generated (*self-conscious*: negative binomial regression, N = 148 pairs, *b* = –.03, SE = .05, *z* = –.58, *p* = .559; *comfortable*: negative binomial regression, N = 148 pairs, *b* = –.01, SE = .04, *z* = –.26, *p* = .798*; dominant*: negative binomial regression, N = 148 pairs, *b* = –.03, SE = .05, *z* = –.47*, p* = .637; *confident*: negative binomial regression, N = 148 pairs, *b* = .01, SE = .04, *z* = .28*, p* = .783). Most importantly, controlling for these measures did not significantly attenuate the effect of modality on idea generation (negative binomial regression, N = 148 pairs, *b* = .18, SE = .07*, z* = 2.46, *p* = .014).

These results suggest that differences in social behaviors do not underly our effect.

G. Linguistic Category Analysis

Method

We used a transcription service to transcribe 109 audio recordings in the first batch of data collection and 149 video recordings in second batch of data collection in the lab (Forty-one groups the first batch of data collection in the lab were not transcribed because of technological issues with the audio recorder. Switching to video in the lab study's second data collection batch reduced these technological errors, and only two groups were not transcribed in this study). We cleaned the transcripts of each group by removing any conversation with the experimenter and then separating the transcripts into two sessions (the 5-minute idea generation session and the subsequent 1-minute idea selection session). We submitted the cleaned transcripts to LIWC2015 (Linguistic Inquiry and Word Count, version 1.6) to determine if there were any language markers that differentiated communication modality for either the generation or selection task. Specifically, LIWC provided the percentage of words used by each participant in 80 word-categories. We then averaged the scores of the two participants to get the average language use per pair in each category and examined if pairs who communicated over video differed from pairs who communicated in person in any of the 80 identified word categories of LIWC. For ease of interpretation and increased power, we collapsed across stimuli.

Results

To ensure that our data meet the assumptions of a linear regression, we tested the homogeneity of variance using the Levene's test. The Levene's test revealed that for five out of the 80 categories, the variances differed by modality (*word count*: $F(1, 514) = 8.41$, $p = .004$; *words relating to seeing*: $F(1, 514) = 5.04$, $p = .025$; *sad words*: $F(1, 514) = 4.95$, $p = .026$; *quantitative words*: $F(1, 514) = 4.30$, $p = .039$; *internet words*: $F(1, 514) = 4.30$, $p = .039$). For these categories, we ran a non-parametric test. Of note, we did not confirm normality for each measure because the central limit theorem states that, at sufficiently large sample sizes (>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8]. Given that we had over 500 observations, our sample greatly surpasses the 100 sample-size threshold and thus it is unnecessary to check for normality.

The results of the LIWC analyses are summarized in Table 1 below. Notably, we found that virtual and in-person pairs do not significantly differ in the number of words spoken ($M_{virtual}$ = 225, SD = 118, $M_{in\text{-}person}$= 217, SD = 90, Kruskal-Wallis rank sum test, N = 258 pairs, χ^2 (1) = .003, *p* = .958, Cohen's *d* = .09, 95% CI: [–.15, .34]). Further, we found that usage in only two (out of 80) word categories differed by communication modality (*second-person singular*: $M_{virtual}$ = 4.12, SD = 2.18, $M_{in\text{-}person}$= 3.60, SD = 2.18, linear mixed effect regression, N = 516 participants, *b* = .52, SE = .21, *t*(256) = 2.55, *p* = .011, Cohen's *d* = .32, 95% CI: [.07, .57]; *sad words*: $M_{virtual}$ = .16, SD = .38, $M_{in\text{-}person}$= .09, SD = .25, Kruskal-Wallis rank sum test, N = 258 pairs, χ^2 (1) = 5.04, *p* = .025, Cohen's *d* = .26, 95% CI: [.01, .50]).

The two (out of 80) significant differences in word usage could be due to random chance, as you would expect around four (out of 80) categories to be significant with an alpha of .05. Regardless, we next examined whether these word usages correlated with idea generation performance. Second-person singular and sad word usage did not significantly correlate with idea generation performance (negative binomial regression, N = 258 pairs, *second-person singular*: *b* = –.02, SE = .02, *z* = –1.05, *p* = .295; *sad words*: *b* = –.18, SE = .12, *z* = –1.52, *p* = .128). Most importantly, the effect of communication modality on idea generation performance is still significant when controlling for sad and second word singular word usage during idea generation (negative binomial regression, N = 258 pairs, *b* = .12, SE = .06, *z* = 2.07, *p* = .039).

Supplementary Table 1: Word usage (calculated using the Linguistic Inquiry and Word Count Database) by communication modality for the idea generation task.

Language Category	Virtual		In-Person		Difference between conditions				Measure predicting idea generation			
	M (SD)		*M* (SD)		*b*	SE	*t*	*p*	*b*	SE	*z*	*p*
Achievement	0.89	(0.88)	0.8	(0.74)	0.09	0.08	1.13	0.259	-0.16	0.05	-3.37	0.001
Adjectives	3.76	(2.04)	3.59	(1.78)	0.17	0.19	0.9	0.369	-0.03	0.02	-1.75	0.08
Common adverbs	4.81	(2.08)	5.09	(2.29)	-0.27	0.2	-1.4	0.162	-0.07	0.02	-3.71	0
Affect words	5.33	(2.66)	5.23	(2.58)	0.1	0.26	0.37	0.713	0.01	0.01	0.51	0.611
Affiliation	1.62	(1.37)	1.75	(1.27)	-0.13	0.13	-1	0.316	-0.04	0.03	-1.52	0.13
Analytical thinking	43.1	(22.8)	42.1	(22.1)	1.01	2.25	0.45	0.654	0.01	0	4.53	0
Anger	0.14	(0.37)	0.16	(0.36)	-0.02	0.04	-0.53	0.596	0.11	0.09	1.24	0.214
Anxiety	0.22	(0.47)	0.3	(0.54)	-0.08	0.05	-1.45	0.149	0.16	0.06	2.56	0.01
Articles	6.41	(2.63)	6.29	(2.85)	0.12	0.29	0.43	0.67	0.01	0.01	0.86	0.388
Assent	5.32	(3.32)	5.19	(3.39)	0.13	0.3	0.44	0.659	0	0.01	0	0.997
Authentic	20.5	(17.8)	19.3	(17.2)	1.17	1.74	0.67	0.505	-0.01	0	-4.06	0
Auxiliary verbs	10.5	(3.03)	10.6	(3.02)	-0.05	0.3	-0.15	0.882	-0.05	0.01	-4.13	0
Biological Processes	1.54	(1.45)	1.63	(1.25)	-0.09	0.14	-0.66	0.511	0.1	0.02	4.33	0
Body	0.64	(0.87)	0.68	(0.79)	-0.04	0.09	-0.52	0.607	0.16	0.04	4.17	0
Cause	3.2	(2.07)	3.01	(1.67)	0.19	0.18	1.05	0.296	-0.02	0.02	-1.26	0.208
Certainty	1.01	(0.93)	1.06	(0.94)	-0.05	0.08	-0.65	0.515	-0.03	0.04	-0.75	0.456
Clout	65.0	(20.1)	63.1	(18.1)	1.88	1.92	0.98	0.329	0	0	0.54	0.586
Cognitive Processes	15.7	(4.44)	15.4	(4.22)	0.26	0.44	0.59	0.558	-0.04	0.01	-5.66	0
Comparatives	4.92	(2.84)	5.03	(3.06)	-0.11	0.29	-0.37	0.709	0	0.01	0.07	0.946
Conjunctions	5.95	(2.5)	5.77	(2.28)	0.19	0.23	0.81	0.417	-0.06	0.02	-4.03	0
Death	0.02	(0.13)	0.02	(0.11)	0	0.01	-0.39	0.695	0.4	0.31	1.26	0.206
Dictionary words	88.3	(4.58)	87.7	(4.34)	0.59	0.47	1.26	0.21	-0.04	0.01	-6.22	0
Differentiation	3.67	(1.81)	3.65	(1.82)	0.02	0.18	0.1	0.922	-0.08	0.02	-3.95	0
Discrepancies	2.92	(1.91)	2.64	(1.64)	0.28	0.18	1.55	0.122	-0.03	0.02	-1.49	0.137
Core Drives and Needs	4.9	(2.29)	5.03	(2.1)	-0.13	0.22	-0.61	0.543	-0.02	0.02	-1.2	0.229
Family	0.07	(0.23)	0.1	(0.28)	-0.03	0.03	-0.98	0.33	0.02	0.14	0.11	0.911
Feeling	0.73	(0.79)	0.73	(0.74)	0	0.07	-0.05	0.957	0.05	0.05	1.02	0.307
Female referents	0.14	(0.34)	0.14	(0.36)	0	0.03	0.13	0.898	0.12	0.11	1.09	0.275
Future focus	0.81	(0.79)	0.82	(0.8)	-0.01	0.07	-0.17	0.863	-0.14	0.05	-2.71	0.007
Past focus	1.5	(1.36)	1.48	(1.18)	0.03	0.12	0.22	0.828	-0.02	0.03	-0.65	0.513
Present focus	14.6	(3.52)	14.5	(3.66)	0.14	0.37	0.38	0.704	-0.05	0.01	-5.37	0
Friends	0.02	(0.1)	0.03	(0.15)	-0.01	0.01	-0.85	0.396	-0.07	0.3	-0.25	0.806
Function Words	54.4	(6.81)	53.6	(7.38)	0.82	0.76	1.08	0.283	-0.03	0	-6.72	0
Health/illness	0.21	(0.48)	0.23	(0.53)	-0.03	0.05	-0.48	0.63	0.1	0.06	1.68	0.093
Hearing	0.93	(0.94)	0.79	(0.83)	0.14	0.08	1.65	0.101	0.08	0.04	2.02	0.043
Home	0.7	(0.9)	0.84	(1.21)	-0.14	0.12	-1.2	0.231	0.1	0.03	4.04	0
1st pers singular	3.41	(1.82)	3.34	(1.78)	0.07	0.17	0.42	0.678	-0.05	0.02	-2.43	0.015
Informal Speech	9.19	(6.33)	9.55	(7.02)	-0.36	0.71	-0.5	0.618	0.01	0	2.58	0.01
Ingesting	0.67	(0.91)	0.74	(0.9)	-0.07	0.1	-0.74	0.462	0.06	0.04	1.6	0.109
Insight	2.63	(1.51)	2.73	(1.48)	-0.1	0.14	-0.71	0.476	-0.09	0.02	-3.71	0
Interrogatives	1.98	(1.4)	2.01	(1.41)	-0.03	0.13	-0.21	0.83	-0.04	0.03	-1.32	0.187
Impersonal pronouns	9.83	(2.95)	9.82	(2.86)	0.01	0.29	0.04	0.97	-0.03	0.01	-2.54	0.011
Leisure	1.72	(1.33)	1.78	(1.49)	-0.07	0.15	-0.44	0.657	0.03	0.02	1.42	0.155

Male referents	0.05	(0.19)	0.06	(0.22)	-0.01	0.02	-0.61	0.546	-0.35	0.2	-1.75	0.079
Money	0.13	(0.32)	0.13	(0.33)	0	0.03	0.14	0.888	-0.15	0.11	-1.28	0.2
Motion	1.68	(1.21)	1.84	(1.19)	-0.16	0.11	-1.39	0.165	0.01	0.03	0.47	0.64
Negations	1.45	(1.11)	1.66	(1.17)	-0.21	0.11	-1.95	0.052	-0.08	0.03	-2.35	0.019
Negative emotion	0.8	(0.9)	0.83	(0.9)	-0.04	0.09	-0.4	0.686	0.08	0.04	2.07	0.038
Netspeak	0.72	(1.66)	1.06	(2.05)	Kruskal-Wallis: χ^2 (1) = 1.84, *p* = .174				0.02	0.01	3.47	0.001
Nonfluencies	3.66	(4.65)	4.14	(5.14)	-0.48	0.54	-0.88	0.381	0.02	0.01	3.47	0.001
Numbers	1.06	(2.16)	0.96	(1.79)	0.1	0.23	0.44	0.662	-0.04	0.02	-2.27	0.023
Perpetual Processes	2.49	(1.39)	2.34	(1.29)	0.15	0.13	1.15	0.252	0.06	0.03	2.3	0.022
Positive emotion	4.52	(2.48)	4.38	(2.35)	0.14	0.24	0.58	0.56	0	0.01	-0.21	0.836
Power	1.12	(1)	1.1	(0.98)	0.02	0.1	0.19	0.85	-0.02	0.03	-0.45	0.653
Personal pronouns	9.42	(2.69)	9	(2.54)	0.42	0.25	1.64	0.102	-0.05	0.01	-3.95	0
Prepositions	12.1	(3.34)	11.6	(3.44)	0.49	0.34	1.44	0.151	-0.02	0.01	-2.32	0.021
Total pronouns	19.3	(3.96)	18.8	(4.04)	0.43	0.41	1.06	0.29	-0.04	0.01	-4.34	0
Question marks	2.86	(2.17)	3.08	(2.48)	-0.22	0.23	-0.93	0.353	0	0.02	0.13	0.894
Quantifiers	1.46	(1.1)	1.24	(0.9)	Kruskal-Wallis: χ^2 (1) = 3.57, *p* = .059				-0.11	0.04	-2.87	0.004
Relativity	8.11	(3.03)	7.69	(3.09)	0.42	0.3	1.4	0.162	-0.02	0.01	-1.56	0.119
Religion	0.05	(0.21)	0.06	(0.22)	-0.01	0.02	-0.44	0.657	0.65	0.15	4.26	0
Reward focus	1.08	(0.92)	1.05	(1.08)	0.03	0.09	0.31	0.754	0.03	0.04	0.91	0.363
Risk/prevention focus	0.41	(0.64)	0.52	(0.67)	-0.11	0.06	-1.67	0.097	0.12	0.05	2.36	0.018
Sadness	0.16	(0.38)	0.09	(0.25)	Kruskal-Wallis: χ^2 (1) = 5.04, *p* = .025				-0.18	0.12	-1.52	0.128
Seeing	0.79	(0.84)	0.73	(0.67)	Kruskal-Wallis: χ^2 (1) = .04, *p* = .835				0.01	0.05	0.22	0.824
Sexuality	0.04	(0.3)	0.01	(0.08)	0.03	0.02	1.36	0.175	0.11	0.17	0.63	0.53
3rd pers singular	0.17	(0.37)	0.16	(0.38)	0.01	0.03	0.32	0.75	0.05	0.1	0.53	0.597
Words>6 letters	10.3	(3.24)	10.6	(3.5)	-0.37	0.34	-1.08	0.283	0.03	0.01	2.86	0.004
Social Words	7.79	(2.7)	7.56	(2.47)	0.23	0.25	0.92	0.358	-0.03	0.01	-2.32	0.021
Space	4.81	(2.08)	4.39	(2.19)	0.42	0.21	1.97	0.05	-0.02	0.02	-1.34	0.181
Swear words	0.04	(0.2)	0.04	(0.17)	0	0.02	-0.19	0.853	0.27	0.2	1.32	0.187
Tentativeness	4.96	(2.56)	4.84	(2.57)	0.13	0.26	0.5	0.619	-0.05	0.01	-3.83	0
3rd pers plural	0.52	(0.69)	0.58	(0.69)	-0.06	0.06	-0.88	0.382	-0.11	0.06	-1.93	0.054
Time	1.65	(1.14)	1.52	(1.26)	0.12	0.12	1.08	0.281	-0.06	0.03	-2	0.045
Emotional Tone	77.7	(22.6)	75.9	(23.7)	1.8	2.17	0.83	0.408	0	0	-0.68	0.494
Regular verbs	18.3	(3.86)	17.9	(4.09)	0.44	0.4	1.12	0.264	-0.04	0.01	-4.36	0
Word Count	224	118	217	(89.5)	Kruskal-Wallis: χ^2 (1) = .03, *p* = .958				-0.00	0.00	-0.88	0.381
1st pers plural	1.2	(1.28)	1.33	(1.15)	-0.13	0.12	-1.12	0.264	-0.07	0.03	-2.14	0.032
Work	0.86	(0.98)	0.87	(0.93)	-0.01	0.1	-0.08	0.934	0.04	0.03	1.19	0.234
2nd person	4.12	(2.18)	3.6	(1.99)	0.52	0.21	2.55	0.011	-0.02	0.02	-1.05	0.295

Legend: For examples of word categories, visit the LIWC manual. To examine the effect of condition on word category usage, we ran linear mixed effect regressions with word usage as the dependent variable, condition (in-person vs. virtual) as the dependent variable, and pair as a random effect (N = 516 participants). As mentioned in text, to ensure that our data meet the assumptions of a linear regression, we tested the homogeneity of variance using the Levene's test. When thdata fail the homogeneity test, we conduct a non-parametric test (Kruskall-Wallis). We did not test for normality because the central limit theorem states that, at sufficiently large sample sizes (>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8].To examine the relationship between word category usage and idea generation performance, we ran negative binomial regressions with creative idea generation dependent variable and word usage as the dependent variable (N = 258 pairs). We confirmed the use of negative binomial regression in Supplementary Information section A.

H. Mimicry

Mimicry is the spontaneous imitation of interaction partners[15]. Research suggests that the propensity to engage in mimicry is innate and occurs often with friends, romantic partners, and even strangers. Is it possible that videoconferencing inhibits participants' innate instinct to unconsciously mimic their partner because virtual interaction is less "natural"? And if so, how might these changes in mimicry affect the collaborative process of idea generation?

The literature provides conflicting predictions about whether the extent to which partners mimic each other helps or hinders idea generation. On one hand, mimicry can foster liking, empathy, and rapport[15] and can reflect coordination, engagement, and willingness to cooperate[16]. By this view, mimicry could improve group cohesiveness and performance. On the other hand, research has also documented that behavioral mimicry can decrease individual creativity by signaling cooperation and belonging to the in-group rather than individualism and breaking away from the status quo[17]. Thus, in order to test the effect of mimicry in our studies, we explore whether modality of communication affects mimicry and whether mimicry correlates with collaborative idea generation performance. Below, we first examine if our experimental manipulation (in person vs. video communication) affected the extent to which pairs mimicked each other's *language* and *facial expressions*. Then, we explore how these forms of mimicry relate to idea generation. We find no evidence that modality affects either form of mimicry.

Linguistic Mimicry. We identified and quantified two forms of linguistic mimicry. First, we examined *language style matching*, which is defined as "the degree to which two people subtly match each other's speaking style"[16]. The language style matching (LSM) score is operationalized as the similarity between the function-word usage of each person in a pair (i.e., the use of personal pronouns, impersonal pronouns, articles, conjunctions, prepositions, auxiliary verbs, adverbs, negations, and quantifiers). As a result, this score reflects synchrony of language style independent of context or conversational topic. For example, consider these two sentences regarding new uses for a frisbee: “have **you** thought **of like** using **the** frisbee **as a** bowl?” and “what **about a** frisbee bowl where someone eats **from it**?” Although they essentially communicate the same idea, these sentences do not have similar function-word usage and thus exhibit low LSM; however, the two sentences “have **you** thought **of like** using **the** frisbee **as a** bowl?” and “**you** could **like** take **the** frisbee and hang **it** up **as a** piece **of** art” would score higher on LSM. LSM is correlated with relationship stability, group cohesion, and mutual understanding[18], but LSM is uncorrelated with self-report measures of similarity or liking, which suggests that LSM reflects nonconscious verbal coordination[16].

Following prior research, we calculated LSM for each function-word category using the following equation: $\mathrm{LSM}_{\mathrm{function\_word}} = 1 - [(\left| \mathrm{p1}_{\mathrm{function\_word}} - \mathrm{p2}_{\mathrm{function\_word}} \right|)/(\mathrm{p1}_{\mathrm{function\_word}} + \mathrm{p2}_{\mathrm{function\_word}} + 0.0001)]$ (p1 and p2 represent the relative scores of the function word for each participant). We then averaged the LSM scores of all function words to generate a mean LSM score for each group. Scores range from 0 to 1, with higher numbers reflecting higher language style matching. Given that our studies were conducted on a college campus with undergraduate students, it is not surprising that groups exhibited high levels of LSM overall (M = .77, SD = .09).

To test if modality affects language style matching, we ran a regression with the standardized LSM score as the dependent variable, study number as a covariate, and modality (in-person vs. virtual) as the independent variable. Importantly, LSM did not significantly differ by modality ($M_{virtual}$ = .77, SD = .10, $M_{in\text{-}person}$ = .77, SD = .08, linear regression, N = 258 pairs, b = .004, SE = .01, t(255) = .36, p = .720, Cohen’s d = .04, 95% CI: [–.20, .29]). As a robustness check, we ran another regression controlling for word count, as verbosity might affect the extent of language style matching. Even when controlling for word count, modality did not significantly affect the LSM score (linear regression, N = 258 pairs, *b* = .01, SE = .01, *t*(254) = .79, *p* = .430). Finally, although LSM's variance did not significantly differ by condition (Levene's test: *F*(1, 256) = 2.58, *p* = .110), the distribution of LSM is significantly non-normal (Shapiro-Wilk normality test: *W* = .90, *p* <.001). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, a non-parametric test which does not assume a type of distribution, reaches the same statistical conclusion (Kruskal-Wallis rank sum test, N = 258 pairs, χ^2 (1) = .001, *p* = .981). Thus, overall, we find no evidence that linguistic mimicry is affected by the modality of communication.

Lastly, to examine if LSM relates to idea generation, we ran a negative binomial regression with number of creative ideas as the dependent variable, study as a covariate, and condition and LSM as predictors. Interestingly, LSM negatively correlates with creative performance (negative binomial regression, N = 258 pairs, *b* = –.63, SE = .29, *z* = –2.18, *p* = .029); however, central to our hypotheses, the effect of modality on creative performance remains significant when controlling for LSM (negative binomial regression, N = 258 pairs, *b* = .13, SE = .05, *z* = 2.46, *p* = .014).

These findings suggest unconscious language style matching does not underly our effect. However, LSM examines linguistic mimicry using functional words that are independent of context (e.g., pronouns, articles, etc.) to investigate uncontaminated language style. In other words, the language style matching index calculated for each pair was independent of the kinds of ideas each partner had. However, some might consider this a conservative measure of linguistic mimicry. Thus, to comprehensively test the effect of modality on linguistic mimicry, we also calculated the degree of similarity between partners across 25 semantic categories provided by LIWC (personal pronouns, first-person pronouns, articles, conjunctions, prepositions, auxiliary verbs, adverbs, negating words, quantitative words, verbs, affect-related words, positive emotion words, negative emotion words, cognitive processing words, tentative words, certain words, affiliation words, assent-related words, nonfluencies, swear words, and informal words) as well as LIWC's overall score on four constructs (Analytic, Clout, Authentic, and Tone, see LIWC dictionary for more information on these constructs and above semantic categories).

Specifically, we gathered the semantic categories from LIWC, generated a vector containing all the semantic categories for each partner, and calculated the distance between these vectors in multidimensional space using cosine similarity[19]. This measure captures "language congruence" across all LIWC categories, with larger scores reflecting higher similarity. Of note, this measure correlates positively with the LSM measure above (Pearson's correlation, N = 258 pairs, r = .30, *p* <.001). As before, to test if modality affects language congruence, we ran a regression with standardized language congruence as the dependent variable, study number as a covariate, and modality as the independent variable. We again found no evidence that modality affects language congruence of partners ($M_{virtual}$ = .91, SD = .08, $M_{in\text{-}person}$ = .92, SD = .07, linear regression, N = 258 pairs, *b* = .01, SE = .01, *t*(256) = .93, *p* = .354), even when controlling for verbosity (linear regression, N = 258 pairs, *b* = .01 , SE = .01 , *t*(254) = 1.23, *p* = .220). Finally, as with LSM, language congruence variance did not significantly differ by condition (Levene's test: $F(1, 256) = .60$, $p = .441$), but the distribution is significantly non-normal (Shapiro-Wilk normality test: $W = .84$, p <.001). Although this is not of great concern (as mentioned above), we ran a non-parametric test as a robustness check. The results provide the same statistical conclusion (Kruskal-Wallis rank sum test, N = 258 pairs, χ^2 (1) = .66, $p = .415$).

Interestingly, unlike LSM, we found that linguistic congruence across 25 semantic categories was not significantly associated with creative idea generation (negative binomial regression, N = 258 pairs, $b = .19$, SE = .37, $z = .53$, $p = .600$). However, matching the LSM results, controlling for linguistic congruence did not significantly attenuate the effect of modality on creative performance (negative binomial regression, N = 258 pairs, $b = .13$, SE = .06, $z = 2.36$, $p = .018$). These results suggest that while mimicry of language style matching potentially hinders idea generation, there is no evidence that using similar words in general to describe ideas positively or negatively relates to idea generation. More importantly, across these analyses, we find no evidence that linguistic mimicry underlies the effect of modality on creative idea generation.

Facial Mimicry. To quantify facial mimicry, we used the OpenFace software to extract the intensity (from 0 to 1) of 18 facial action units (e.g., nose wrinkle, lip tightener, cheek raiser[20]) from each frame of each video of participants recorded in the second batch of data collection in the lab. We then averaged the facial action intensities for each participant, generated a vector containing the mean facial action intensities for each partner, and calculated the distance between these vectors in multidimensional space using cosine similarity. As with the linguistic congruence measure above, higher scores reflect more similarity in facial expressions. To test if modality affects facial expression congruence, we ran a regression with facial expression congruence as the dependent variable and modality as the independent variable.

As before, we found that modality did not significantly affect facial congruence of partners ($M_{virtual}$ = .86, SD = .08, $M_{in\text{-}person}$ = .85, SD = .09, linear regression, N = 145 pairs, *b* = .01, SE = .01 , *t*(143) = .40, *p* = .688). As with the other measures of congruence, facial congruence variance did not significantly differ by condition (Levene's test: *F*(1, 143) = .87, *p* = .352), but the distribution was significantly non-normal (Shapiro-Wilk normality test: *W* = .91, *p* <.001). Although this is not of great concern (as mentioned above), we ran a non-parametric test as a robustness check. The results provide the same statistical conclusion (Kruskal-Wallis rank sum test, N = 145 pairs, χ^2 (1) = .01, *p* = .921).

Next, we examined whether facial congruence was associated with idea generation performance. We found that facial congruence does not significantly relate to creative idea generation (negative binomial regression, N = 145 pairs, *b* = –.72, SE = .40, *z* = –1.77, *p* = .076), and controlling for facial congruence did not significantly attenuate the effect of modality on creative performance (negative binomial regression, N = 145 pairs, *b* = .16, SE = .07, *z* = –2.17, *p* = .030).

In sum, using three different operationalizations of mimicry (language style matching, linguistic congruence, and facial expression congruence), we find no evidence that modality of communication affected linguistic mimicry in our studies.

I. Communication Coordination

This section covers the methods and analyses used to examine the effect of modality on communication coordination. We first examined the effect of condition on self-report measures collected via surveys at the end of the lab study. Specifically, in the first batch of data collection, we asked each participant to indicate "How often did you feel like you were talking over each other?" from 1 = *very rarely* to 7 = *very often*. To ensure that this measure meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. The variances between modalities were not significantly different from each other ($F(1, 296) = 1.60$, $p = .208$), however the distribution is significantly non-normal (Shapiro-Wilk normality test: $W = .72$, $p <.001$). This result should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we report the result of the non-parametric test as well as the linear regression model.

We find that virtual participants self-report experiencing more communication coordination friction than in-person pairs ($M_{virtual} = 2.15$, SD = 1.47, $M_{in\text{-}person} = 1.83$, SD = 1.31, mixed effect linear regression, N = 298 participants, $b = .32$, SE = .16, $t(296) = 2.00$, $p =.047$, Cohen's $d = .23$, 95% CI [–.00, .46]). A non-parametric test also reveals a significant difference by conditions (Kruskal-Wallis rank sum test, N = 150 pairs, χ^2 (1) = 5.42, $p = .020$).

To follow-up on this result, we quantified communication coordination friction via three complementary transcription metrics: the number of words spoken, the number of times the transcriber noted "crosstalk" during the interaction (which reflects when two people are speaking over each other) and the number of speaker switches ("back-and-forths") each pair exhibited in their transcripts. These metrics capture different outcomes of communication coordination. Pairs who can seamlessly determine when each partner speaks should speak more words, switch back and forth between speakers more fluidly during conversation, and engage in less crosstalk. To ensure that these measures meet the assumptions of a linear regression, we evaluated whether each measure's variances were homogenous by conducting the Levene's test[5] and whether their distributions were normal by conducting the Shapiro-Wilk normality test[6]. All measures were significantly non-normal (Shapiro-Wilk normality test: *back-and-forth*: $W = .98$, $p = .02$; *crosstalk*: $W = .61$, $p < .001$; *word count:* $W = .94$, $p < .001$). Further, although the back-and-forth measure did not significantly violate the homogeneity of variance assumption (Levene's test: $F(1, 147) = 1.13$, $p = .289$), the crosstalk measure and word count variances were significantly different by condition (Levene's test, *crosstalk*: $F(1, 147) = 10.15$, $p = .002$; w*ord count*: $F(1, 514) = 8.41$, $p = .004$). Thus, we will only report the non-parametric test for crosstalk and word count, and we will report both the linear regression model and the non-parametric test for the back-and-forth measure.

In contrast with participants' subjective perceptions, we found that virtual and in-person pairs do not significantly differ in the number of words spoken ($M_{virtual} = 225$, SD = 118, $M_{in\text{-}person} = 217$, SD = 90, Kruskal-Wallis rank sum test, N = 258 pairs, $\chi^2 (1) = .003$, $p = .958$, Cohen's $d = .09$, 95% CI: [–.15, .34]). Further, virtual pairs also engaged in significantly *less* crosstalk ($M_{virtual} = .65$, SD = 1.12, $M_{in\text{-}person} = 1.60$, SD = 2.40, Kruskal-Wallis rank sum test, N = 149 pairs, $\chi^2 (1) = 8.12$, $p = .004$). However, in line with participants' subjective perceptions, virtual groups did engage in significantly *fewer* speaker switches than in-person groups ($M_{virtual}$ =

50.1, SD = 19.3, $M_{in\text{-}person}$ = 59.6, SD = 21.1, linear regression, N = 149 pairs, b = 9.48, SE = 3.31, t(147) = 2.87, p =.005). A non-parametric test revealed similar results (Kruskal-Wallis rank sum test, N = 149 pairs, χ^2 (1) = 6.81, p = .009). Perhaps virtual groups engage in more deliberate turn-taking due to perceived challenges in conversation coordination during virtual interaction or due to norms of virtual interaction. Regardless, most importantly, when controlling for crosstalk and speaker switches, the effect of modality on number of creative ideas generated remains significant (negative binomial regression, N = 149 pairs, b = .16, SE = .07, z = 2.21, p =.027).

These results suggest that it is unlikely that communication coordination fully explains the negative effect of virtual interaction on idea generation.

J. Examining Team Processes Identified in Prior Literature

Prior work has identified multiple important team processes underlying idea generation performance. Below, we explore whether communication modality alters previously identified social and cognitive team processes and whether controlling for these team processes meaningfully attenuates our documented negative effect of virtual collaboration on idea generation.

Social Processes

Many social processes, such as fear of evaluation (and resulting self-censorship), social facilitation, social loafing, social sensitivity, dominance, and perceptions of team performance, unfold during and affect group idea generation. Does modality alter these social processes, and could that explain the idea generation effect we observe? On one hand, “virtual” group members might have reduced social presence[21]: an individual on a video call is typically viewed on a screen that is smaller than their real-life counterpart, and the camera angle on video calls usually omits the lower body from view. Thus, the innate social processes that underly team performance could be dampened (for better or for worse) during virtual interaction due to lower perceived social presence. For example, it may be less stressful to speak up in a group during videoconferencing.

On the other hand, eye gaze data from second batch of data collection in the lab finds that participants look at their partner for a larger proportion of the time when interacting virtually. This increased attention to one’s partner could increase the awareness of others’ presence and heighten the influence of interpersonal factors. For example, looking at one’s partner more when interacting virtually could increase fear of evaluation and resulting self-censorship. Thus, prior literature supports the notions that video interaction could either dampen or heighten social processes. Below, we empirically examine these possibilities across seven unique social factors identified by prior research.

Social Sensitivity. Prior research has found that the social sensitivity of groups—that is, group members' ability to read and get a sense of their other group members— improves group performance across a wide range of tasks, including solving puzzles, making moral judgments, idea generation, and negotiating[22]. In this prior research, social sensitivity was measured by averaging participants' scores on the "Reading the Mind in the Eyes" test, in which participants indicate the expressed emotion in a series of photographed eyes[23]. Could virtual interaction hinder participants' ability to read their partner, and if so, could this decrease in social sensitivity explain the negative effect of virtual interaction on idea generation? We suspected this was not the case because virtual pairs viewed a large display of their partners' face and interacted in real time, which should, in theory, allow them to gauge their partner's reactions. In fact, the virtual interaction display is not unlike the common test of social sensitivity used in the experiments mentioned earlier, the "Reading the Mind in the Eyes" test, where participants read emotions from images of eyes on a screen. On the other hand, virtual interaction is certainty less natural than in-person communication and could inhibit participants' innate social ability to get a sense of their partner. Thus, we examined if participants' ability to read their partner differed based on communication modality and whether this could explain the observed difference in idea generation by communication modality.

To test this, we leveraged the personality data collected at the end of the second batch of data collection in the lab, in which participants each indicated their perceptions of their partner's personality and then rated their own personality (Big 5: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) using the ten item personality inventory (TIPI)[24].

We captured social sensitivity by measuring the proximity of a person's perception of their partner (after 15 minutes of interaction) to the partner's self-views[25]. If virtual interaction hindered participants' ability to get a sense of their partner, these participants should less accurately estimate their partner's personality. Specifically, we calculated the congruence between one's perceptions of their partner's personality and the partner's self-reported personality across all five personality dimensions using cosine similarity (the same approach as the mimicry analyses above). By doing this, we are able to quantify the extent to which someone's perceptions of their partner match the partner's self-reported personality on a continuum: higher values suggest greater social sensitivity (i.e., more congruence between an individual's sense of their partner and their partner's self-reported personality) and lower values suggest lower social sensitivity (i.e., less congruence between an individual's sense of their partner and their partner's self-reported personality).

To ensure that our social sensitivity measure meets the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. The social sensitivity measure does not violate the homogeneity of variance assumption, as the variances between modalities were not significantly different from each other (Levene's test: $F(1, 296) = 2.63$, $p = .106$), but the distributions was significantly non-normal (Shapiro-Wilk normality test: W = .97, $p <.001$). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample size (N > 100)[8]. However, in the spirit of comprehensiveness, we report results from both a linear regression model and a non-parametric test (which makes no assumptions regarding the underlying distribution) below.

To test if modality affected social sensitivity, we ran a linear mixed effects regression with group number as a random effect, congruence between an individual's perception of their partner and partner's self-view as the dependent variable, and modality as the independent variable. There was no significant difference in social sensitivity by modality ($M_{virtual}$ = .42, SD = .26; $M_{in\text{-}person}$ = .41, SD = .24, linear mixed effect regression, N = 298 participants, *b* = .02, SE = .03, *t*(147) =.56, *p* = .574, Cohen's *d* = .07, 95% CI [–.16, .30]). A non-parametric test reaches the same statistical conclusion (Kruskal-Wallis rank sum test, N = 151 pairs, χ^2 (1) = .26, *p* = .611).

Next, we examined the relationship between social sensitivity and idea generation. Social sensitivity did not significantly relate to idea generation (negative binomial regression, N = 149 pairs, *b* = .27, SE = .19, *z* = 1.39, *p* =.164). Further, and most importantly, controlling for social sensitivity did not significantly attenuate the effect of modality on creative performance (negative binomial regression, N = 149 pairs, *b* = .15, SE = .07, *z* = 2.14, *p* =.033). Thus, a participant's ability to read their partner cannot explain the effect of virtual interaction on creative performance.

Fear of Evaluation. Research suggests that evaluation apprehension can decrease idea generation performance through self-censorship[26]. In other words, idea generation can be hampered when a participant withholds a very creative but "out-there" idea out of fear of being judged. Could virtual communication hinder idea generation by focusing individuals' gaze on their partner and heightening evaluation apprehension?

We examined this potential alternative account using survey data collected in the lab study. First, in the first batch of data collection in the lab, we directly asked participants how much they engaged in self-censorship due to evaluation apprehension (i.e., "*How often did you not say an idea because you were worried about what your partner would say about it?"*) and their perceptions of criticism ("*How often did your partner criticize your ideas?*" and "*How often did you criticize your partner's ideas*," α = .85). Second, in the second batch of data collection in the lab, we asked participants about their self-presentation concerns (i.e., "*In this experiment, I have been somewhat concerned about the way I've presented myself to my partner"* and *"In this experiment, I have been concerned about what my partner might think of me.",* α = .81). Lastly, in the second batch of data collection in the lab, we indirectly measured concerns of criticism by measuring perceptions of partner warmth (*How tolerant, warm, good-natured, and sincere is this person?*).

To ensure that these measures meet the assumptions of a linear regression, we evaluated whether the variance for our measure of average creativity was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. Our models do not violate the homogeneity of variance assumption: the variances between modalities were not significantly different from each other for any of the measures (Levene's test: *self-censorship*: $F(1, 296) = .18$, $p = .669$; *perceptions of criticism*: $F(1, 296) = .01$, $p = .914$; *perceived partner warmth*: $F(1, 297) = .96$, $p = .328$; *self-presentation concerns*: $F(1, 297) = .13$, $p = .717$). However, all four measures are significantly non-normal (Shapiro-Wilk normality test for *self-censorship*: $W = .58$, $p < .001$; *perceptions of criticism*: $W = .63$, $p < .001$; *perceived partner warmth*: $W = .92$, $p < .001$; and *self-presentation concerns*: $W =$

.95, *p* <.001). Although this is not of great concern (as mentioned above), as a robustness check, we will also report the results of a non-parametric test below.

We found that none of these measures approximating evaluation apprehension significantly differed by condition. First, participants reported similar levels of self-censorship due to evaluation apprehension ($M_{virtual}$ = 1.57, SD = 1.12, $M_{in\text{-}person}$ = 1.52, SD = 1.02, mixed effect linear regression, N = 298 participants, *b* = .05 SE = .13, *t*(148) = .42, *p* =.676, Cohen's *d* = .05, 95% CI [–.18, .28]) and similar perceptions of criticism ($M_{virtual}$ = 1.48, SD = .87, $M_{in\text{-}person}$ = 1.47, SD = .78, mixed effect linear regression, N = 298 participants, *b* = .01 SE = .10, *t*(148) = .11, *p* =.917, Cohen's *d* = .01, 95% CI [–.21, .24]) in each modality condition. Further, modality did not significantly affect perceived partner warmth ($M_{virtual}$ = 5.89, SD = .93, $M_{in\text{-}person}$ = 5.97, SD = .86, mixed effect linear regression, N = 299 participants, *b* = .08, SE = .11, *t*(148) = .72, *p* =.475, Cohen's *d* = .09, 95% CI [–.14, .31]) or self-presentation concerns ($M_{virtual}$ = 2.83, SD = 1.09, $M_{in\text{-}person}$ = 2.58, SD = 1.04, mixed effect linear regression, N = 299 participants, *b* = .25, SE = .13, *t*(148) = 1.90, *p* =.060, Cohen's *d* = .23, 95% CI [–.004, .46]). Non-parametric tests (which do not assume a type of distribution) reach the same statistical conclusion (Kruskal-Wallis rank sum test, N = 150 pairs, *self-censorship*: χ^2 (1) = .17, *p* = .677; *perceptions of criticism*: χ^2 (1) = .01, *p* = .921; Kruskal-Wallis rank sum test, N = 150 pairs, *perceived partner warmth*: χ^2 (1) = .22, *p* = .638, *self-presentation concerns*: χ^2 (1) = 3.19, *p* = .074).

We next examined whether these measures correlate with idea generation performance and whether controlling for these measures affects the statistical significance of our effect of modality on creative ideation. Both self-censorship and perceptions of criticism negatively correlated idea generation (negative binomial regression, N = 150 pairs, *self-censorship*: *b* = –.13, SE = .05, *z* = –2.50, *p* = .013, *perceptions of criticism*: *b* = –.17 SE = .06, *z* = –2.81, *p*

=.005). However, controlling for these measures does not significantly attenuate the effect of communication modality on ideation (negative binomial regression, N = 150 pairs, *b* = .16, SE = .07, *z* = 2.38, *p* = .017). Further, perceived partner warmth and self-presentation concerns did not significantly relate to idea generation performance (*perceived partner warmth*: *b* = .01, SE = .05, *z* = .25, *p* = .800; *self-presentation concerns*: *b* = –.002, SE = .04, *z* = –.06, *p* = .953). Thus, it is not surprising that controlling for these measures does not significantly attenuate the effect of communication modality on ideation (negative binomial regression, N = 150 pairs, *b* = .15, SE = .07, *z* = 2.13, *p* = .034) Thus, it seems unlikely evaluation concerns underly the negative effect of virtual communication on idea generation.

Enjoyment and Illusion of Productivity. Prior research finds that face-to-face groups are more satisfied with their performance during idea generation tasks than people working alone, although these differences often do not translate to objective performance[27]. Could interacting over video reduce this feeling of satisfaction, and could this explain the decrease in creative idea generation among virtual groups?

To test this, in the lab study, we measured both enjoyment of idea generation ("*How much did you enjoy generating ideas in the idea generation task?*") and satisfaction with performance ("*How satisfied are you with your group's performance in the idea generation task?*") at the end of the study. To ensure these measures meet the assumptions of a linear regression model, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5]. This measure does not significantly violate the homogeneity of variance assumption: the variances of each measure did not significantly differ by condition (*satisfaction*: *F*(1, 597) = .33, *p* = .567; *enjoy*: *F*(1, 597) = .02 *p* = .897). Of note, we did not confirm normality for each measure because the central limit theorem states that, at sufficiently large sample sizes

(N>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8]. Given that we had over 500 observations, our sample greatly surpasses the 100 sample-size threshold and thus it is unnecessary to check for normality.

We found that there was no evidence of differences by communication modality for creative task enjoyment ($M_{virtual}$ = 5.05, SD = 1.37, $M_{in\text{-}person}$ = 5.16, SD = 1.34, mixed effects linear regression, N = 599 participants, *b* = .10, SE = .11, *t*(367) = .87, *p* = .387, Cohen's *d* = .076, 95% CI [–.085, .236]) or satisfaction with performance ($M_{virtual}$ = 4.84, SD = . 1.31, $M_{in\text{-}person}$ = 4.79, SD = 1.31, mixed effects linear regression, N = 599 participants, *b* = .04, SE = .11, *t*(350) = .39, *p* = .700, Cohen's *d* = .036, 95% CI [–.197, .124]). Unsurprisingly, the negative effect of communicating virtually on idea generation held even when controlling for these metrics (negative binomial regression, N = 300 pairs, *b* = .17, SE = .05, *z* = 3.25, *p* = .001). Thus, we find no evidence that perceptions of performance or enjoyability of task explain our documented effect.

Dominance. Prior work has demonstrated that group hierarchy and dominance can affect creativity—specifically, people of higher status or power (e.g., White or older group members) can dominate discussion and prevent opposing ideas from being considered or perhaps from being voiced at all[28]. Could the increased visual focus on one's partner when interacting over video increase the salience of status cues, such as posture or size, and as a result, increase the likelihood that one participant dominates the conversation?

We tested this proposition in three different ways. First, we objectively captured conversation dominance by subtracting the word count of each individual partner and dividing this by the total word count. If participants spoke equal numbers of words, this metric would yield a score of 0, representing no verbal dominance. If one participant spoke the entire time, this metric would yield a score of 1, representing total verbal dominance. Interestingly, we found that the verbal dominance was higher for virtual pairs than in-person pairs ($M_{virtual}$ = .26, SD = .17, $M_{in\text{-}person}$ = .21, SD = .15, linear regression, N = 258 pairs, *b* = .05, SE = .02, *t*(256) = 2.57, *p* = .011, Cohen's *d* = .32, 95% CI [–.07, .57]). However, in this context, dominance did not significantly relate to idea generation (negative binomial regression, 258 pairs, *b* = –.17, SE = .17, *z* = –.95, *p* =.341), and controlling for dominance did not significantly attenuate the effect of communication modality on creative idea generation (negative binomial regression, 258 pairs, *b* = .13, SE = .06, *z* = 2.26, *p* =.024).

Next, we asked condition-blind observers to watch a 30 second video of each participant and rate how dominant that participant seemed (see the Supplementary Information section F for more detail). Providing convergent validity for the verbal dominance metric calculated via word count, observers rated virtual partners as more dominant (*M* = 3.45, SD = 1.68) than in-person partners (*M* = 3.08, SD = 1.63, linear mixed effect regression, N = 1380 ratings, *b* = .40, SE = .11, *t*(130) = 3.56, *p* < .001, Cohen's *d* = .22, 95% CI [.11, .33]). However, as before, we find no evidence that dominance relates to idea generation (negative binomial regression, N = 148 pairs, *b* = –.02, SE = .05, *z* = –.47, *p* =.637), and controlling for dominance did not significantly attenuate the effect of communication modality on creative idea generation (*b* = .17, SE = .07, *z* = 2.28, *p* =.023). Thus, using two objective measures of dominance (conversational dominance and observed dominance), we find that virtual groups exhibit higher levels of dominance. As mentioned earlier, this may be due to increased attention spent on their partner, which exacerbated the effect of status cues such as age. However, importantly, this difference in dominance does not explain our idea generation effect.

Since it is possible that the mere feeling of dominance could affect creative output, we also examined subjective perceptions of dominance by asking participants at the end of the second batch of data collection in the lab how dominant they felt during the interaction[29]. We found that there was no evidence of difference in feelings of dominance by condition ($M_{virtual}$ = 54.7, SD = 17.2., $M_{in\text{-}person}$ = 56.0, SD = 17.0, linear mixed effect regression, N = 299 participants, *b* = 1.25, SE = 1.98, *t*(297) = .64, *p* = .526, Cohen's *d* = .073, 95% CI [–.154, .301]), and thus, unsurprisingly, controlling for subjective feelings of dominance did not significantly attenuate the negative effect of virtual interaction on creative idea generation (negative binomial regression, N = 150 pairs, *b* = .15, SE = .07, *z* = 2.14, *p* = .032). We lastly calculated subjective dominance asymmetry in each pair (i.e., the extent to which participants differed in feelings of dominance within a pair) by subtracting the feelings of dominance of each individual partner and dividing this by the total feelings of dominance. Again, subjective dominance asymmetries did not significantly differ by communication modality ($M_{virtual}$ = .20, SD = .20, $M_{in\text{-}person}$ = .18, SD = .15, linear regression, N = 150 pairs, *b* = .01, SE = .03, *t*(147) = .43, *p* = .667, Cohen's *d* = .071, 95% CI [–.253, .395]), and controlling for this asymmetry did not significantly attenuate our documented effect (negative binomial regression, N = 150 pairs, *b* = .14, SE = .07, *z* = 1.99, *p* = .047).

To ensure that each of these measures meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. None of the measures violate the homogeneity of variance assumption, as the variances of each measure did not significantly differ by condition (*verbal dominance*: *F*(1, 256) = .96, *p* = .329; *judges' perceptions of dominance*: *F*(1, 1378) = 3.21, *p* = .074; *subjective dominance*: *F*(1, 297) = .06, *p*

= .808; and *dominance asymmetry*: $F(1, 147) = 1.76$, $p = .187$). However, all four measures were significantly non-normal (Shapiro-Wilk normality test for *verbal dominance*: $W = .94$, $p < .001$; *judges' perceptions of dominance*: $W = .93$, $p < .001$; *subjective dominance*: $W = .97$, $p < .001$; and *dominance asymmetry*: $W = .80$, $p < .001$). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the analyses using non-parametric tests (which do not assume an underlying distribution type). These tests reach the same statistical conclusion (Kruskal-Wallis rank sum test: *verbal dominance*: N = 258 pairs, χ^2 (1) = 5.72, $p = .017$; *judges' perceptions of dominance*: N = 148 pairs, χ^2 (1) = 13.26, $p < .001$; *subjective dominance*: N = 150 pairs, χ^2 (1) = .05, $p = .826$, *dominance asymmetry*: N = 150 pairs, χ^2 (1) = .21, $p = .650$).

Together, although we find that virtual interaction increases the extent to which a participant verbally dominates a conversation, there is nothing to suggest that dominance behaviors underlie the negative effect of virtual communication on creative idea generation.

Social Loafing. When people work together in groups (vs. alone), their contribution to a task becomes more dispensable[30]. Increased dispensability encourages participants to shirk responsibility, and as a result, group performance suffers. Is it possible that the communication modality affected the extent to which participants engaged in social loafing and could this explain the negative effect of virtual interaction on idea generation performance?

To test this, in the first batch of data collection in the lab, we asked participants to estimate the extent to which they and their partner had done their fair share of idea generation.

To ensure that our data meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. The variances of each measure did not significantly differ by condition (Levene's test: *self*: $F(1, 146) = .42$, $p = .518$; *other*: $F(1, 146) = .01$, $p = .910$). However, both measures were significantly non-normal (Shapiro-Wilk normality test for *self*: $W = .95$, $p <.001$; *other*: $W = .93$, $p <.001$). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the analyses using a non-parametric test (which makes no assumptions regarding the underlying distribution) and include those results with the linear regression model results below.

We found no evidence of differences in self- or other-perceptions of social loafing by communication modality (linear mixed effect regression, N = 298 participants, *self*: $M_{virtual}$ = 5.37, SD = 1.28, $M_{in\text{-}person}$ = 5.49, SD = 1.20, $b = .12$, SE = .14, $t(296) = .84$, $p = .404$, Cohen's d = .10, 95% CI [–.13, .32], *other*: $M_{virtual}$ = 5.71, SD = 1.21, $M_{in\text{-}person}$ = 5.90, SD = 1.17, $b = .19$, SE = .14, $t(296) = 1.34$, $p = .180$, Cohen's d = .16, 95% CI [–.07, .38]). Non-parametric tests reach the same statistical conclusion (Kruskal-Wallis rank sum test, N = 149 pairs: *self*: N χ^2 (1) = .38, $p = .54$; *other*: χ^2 (1) = 2.43, $p = .12$).

Next, we examined the relationship between social loafing and idea generation. We found that while self-perceptions of social loafing (“*Did I do my fair share of the idea generation?”*) did not significantly relate to group performance (suggesting lack of self-insight or motivated reasoning[31], negative binomial regression, N = 150 pairs, $b = .03$, SE = .05, $z = .64$, $p = .521$), perceptions of one's partner doing their fair share positively correlated with group idea generation performance (negative binomial regression, N = 150 pairs, $b = .20$, SE = .05, $z = 4.32$, $p <.001$). Given that social loafing did not significantly differ by condition, it was not surprising that controlling for other- and self-perceptions of social loafing did not significantly attenuate our effect (negative binomial regression, N = 150 pairs, $b = .15$, SE = .07, $z = 2.06$, $p = .039$). Thus, there is no evidence that the negative effect of virtual interaction on creative idea generation can be explained by increased social loafing among virtual pairs.

Of note, prior work demonstrates that social loafing increases when group size increases[30]. To our knowledge, no one has explored whether group size similarly increases social loafing during video interaction. We were able to test this in our virtual-only study (described in detail in Supplement G). We found that, replicating in-person studies, groups of four reported more social loafing than groups of two (reverse coded; higher values suggest less social loafing, see Extended Data Table 5 for details, $M_{2\text{-}person}$= 4.61, SD = .73, $M_{4\text{-}person}$ = 4.25, SD = 1.07, Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = 10.28, $p = .001$, Cohen's $d = –.38$, 95% CI [–.65, –.11]). This provides further evidence that social loafing operates similarly in video contexts.

Social Facilitation. Prior work demonstrates that the presence of others can be arousing and that this increased arousal can enhance performance[1]. Could virtual interaction decrease arousal due to lower social presence, and could decreased arousal explain why virtual interaction negatively affects idea generation?

To test this, in the second batch of data collection in the lab, we directly asked participants to indicate their arousal on a validated and popular one-item pictorial scale ranging from 0 to 100[29]. To ensure that our data meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. The variances of the arousal measure did not significantly differ by condition (Levene's test: $F(1, 297) = .54$, $p = .463$), but the distribution was significantly non-normal (Shapiro-Wilk normality test: $W = .97$, $p < .001$). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Further, according to the central limit theorem, even with skewed data, the sample means still approximate a normal distribution with our sample sizes (N > 100)[8]. Nevertheless, as a robustness test, we re-ran the analyses using a non-parametric test (which makes no assumptions regarding the underlying distribution) and include those results with the linear regression model results below.

We found there was no significant difference in self-reported arousal by condition (linear mixed effect regression, N = 299 participants, $M_{virtual} = 46.2$, SD = 20.6, $M_{in\text{-}person} = 41.8$, SD = 20.2, $b = 4.41$, SE = 2.47, $t(148) = 1.79$, $p = .076$, Cohen's $d = .22$, 95% CI [–.01, .45]). This was replicated using a non-parametric test (Kruskal-Wallis rank sum test, N = 299 participants, χ^2 (1) = 2.94, $p = .086$). Further, this measure of arousal did not significantly relate to idea generation (negative binomial regression, N = 150 pairs, $b = .002$, SE = .002, $z = .80$, $p = .426$), and, thus, unsurprisingly, controlling for arousal did not significantly attenuate the negative effect of virtual interaction on idea generation (negative binomial regression, N = 150 pairs, $b = .16$, SE = .07, $z = 2.26$, $p = .024$).

Thus, although there is a slight (nonsignificant) effect of communication modality on arousal, this difference did not explain our effect. This is consistent with prior research which finds that the effect size of social facilitation is generally quite small[33].

Cognitive Process: Production Blocking

In addition to social processes, prior research finds that because only one person can speak at time, interactive groups prevent cognitive expansion through “production blocking.” Production blocking occurs when someone does not voice an idea either because they forgot while waiting for their turn to speak or because the conversation went elsewhere and they felt it was no longer appropriate[26,34]. Theoretically, production blocking should not differ between conditions because participants in both conditions shared “airtime” by generating ideas by speaking out loud. Nevertheless, we still wanted to account for this important cognitive team process in our analyses and examine if production blocking could explain our documented effect.

To test this, after the tasks, we asked participants in the first batch of data collection in the lab, “*Were you able to voice all of your ideas?”* from 1 = *not at all* to 7 = *very much*. To ensure that this measure meets the assumptions of a linear regression, we evaluated whether the data’s variance was homogenous by conducting the Levene’s test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. The variances did not significantly differ by condition (Levene’s test: $F(1, 296) = 3.50$, $p = .063$), but the distribution was significantly non-normal (Shapiro-Wilk normality test: W = .72, $p <.001$). As mentioned multiple times in these supplemental analyses, this is not of great concern, but as a robustness check, we also report the results of a non-parametric test below.

We found virtual pairs did not experience significantly more production blocking than in-person pairs (i.e., were less able to voice all of their ideas, $M_{virtual} = 6.06$, SD = 1.23, $M_{in\text{-}person} = 6.32$, SD = 1.02, linear mixed effect regression, N = 298 participants, $b = .26$, SE = .14, $t(148) = 1.88$, $p = .062$, Cohen’s $d = .23$, 95% CI [.00, .46]). A non-parametric test supports this conclusion (Kruskal-Wallis rank sum test, N = 150 pairs, χ^2 (1) = 1.77, $p = .183$). Further, and more importantly, controlling for production blocking did not significantly attenuate the negative effect of virtual interaction on creative idea generation (negative binomial regression, N = 150 pairs, *b* = .15, SE = .07, *z* = 2.10, *p* = .036).

Summary

At first blush, it may be surprising that the effect of virtual interaction on creative idea generation is not explained by these interpersonal processes. We speculate that this is because, for the type of task in our studies, videoconferencing sufficiently mimics many of the social aspects of in-person interaction—assuming good network connectivity, participants can see and hear each other's faces and voices in real-time and even get at least a limited sense of their partners' surroundings. In contrast to our context, it is possible that interpersonal processes play a more important role during videoconferencing interactions with explicit social goals, such as getting to know someone. How might virtual interaction affect these interactions? We leave this question to future research.

K. Group Size Survey

122 engineers from a large, multinational telecommunication infrastructure company voluntarily participated in a survey about their work-from-home behavior during the pandemic. In this survey, engineers first reported the proportion of video calls in a typical week that involved working on creative tasks, which were defined as "problem-solving, idea generation, and idea development." The response options were, "1 = *almost all of my meetings DO NOT involve working on creative tasks*", "2 = *most of my meetings do NOT involve working on creative tasks"*, "3 = *a majority of my meetings do NOT involve working on creative tasks*", "4 = *around 50% of my meetings involve working on creative tasks*", "5 = *a majority of my meetings involve working on creative tasks*", "6 = *most of my meetings involve working on creative tasks*" and "7 = *almost all of my meetings involve working on creative tasks*." Engineers then indicated the percent of video calls involving creative tasks that were one-on-one (a total of 2 participants on a call), from 0 to 100% in 10% increments. Lastly, engineers responded to two exploratory items: first, how many people, on average, were on video calls that included more than one other person (i.e., have three or more people total) for creative tasks (from *3 people* to *8+ people*), and second, how many people they preferred to have on a video call (from 1 = *strongly prefer one-on-one* to 7 = *strongly prefer 3 or more people*).

First, supporting the validity of this sample, 59% of the engineers reported that at least half of their meetings involved creative tasks. Second, supporting the relevance of our experimental context of two-person idea generation, the engineers reported that, on a typical week, 41% (SD = 25%) of meetings involving creative tasks were one-on-one. Interestingly, engineers reported slight preference for 3 or more people on the video call (*M* = 4.54, SD = 1.70), although 47% of engineers either preferred one-on-one or had no preference. Lastly, among larger groups, the average number of people on the group call was 4.75 (SD = 1.27). These data demonstrate that the proportion of one-on-one meetings in the context of idea generation is substantial and that our context is relevant to real-world collaborative creativity.

L. Established Teams

Importantly, many in-person and virtual creative teams work together for months or years and, as a result, their team performance is subject to higher-order factors, such as hierarchy, specialization, and idiosyncratic group norms. Prior work has examined how virtual teams navigate these processes relative to in-person teams[35]. Our one-shot, two-person context likely renders these higher-order factors negligible, which enabled us to specifically focus on how communication modality shapes the cognitive processes underlying idea generation and selection. Would our effects generalize to established teams? Our field studies provide some insight into this possibility, as each session was conducted within an organic team (ranging from 16 to 112 engineers). Within these organic teams, we randomly assigned engineers to work together in pairs of two. Due to this procedure, it is possible that an engineer would be paired with a friend or a familiar colleague. However, this chance varies by the size of the organic team. If you are on a small team, you are more likely to be familiar with a larger proportion of the team, thus decreasing the chance that you would be randomly paired with someone you haven't worked with before. If you are on a larger team, you likely haven't interacted closely with everyone on the team, thus increasing the chance that you would be randomly paired with someone you hadn't worked with before. In other words, as the size of the organic team grows, so does the chance that someone is paired with someone they are unfamiliar with.

To explore whether the effect of modality generalizes to teams who work together often, we examined whether or not the effect of modality differs by the size of the organic team recruited to be in each session. If the negative effect of virtual interaction reduces among people who work together, then smaller organic teams should observe an attenuated effect of virtual interaction, as many engineers were likely paired with a person they had closely interacted with before.

Teams varied substantially in size. The mean session size was 64 engineers, and the sessions sizes ranged from 16 engineers to 112 engineers. We ran a linear mixed effect regression, with session as a random effect, country as a covariate, number of ideas generated as the dependent variable, and communication modality (in person vs. virtual), organic team size, and their interaction as predictors. (We chose this model because the negative binomial model was "nearly unidentifiable," and linear mixed effect regression is a suitable alternative). There was no significant interaction between communication modality and team size (linear mixed effect regression, N = 745 teams, *b* = .01, SE = .02, *t*(723) = .65, *p* = .518). Together, these data provide initial evidence that the negative effect of virtual interaction persists even among people who know each other and work together often. However, we believe that a future examination of how modality affects the impact of higher-order factors, such as hierarchy, on idea generation would be worthwhile.

M. Group Size Lab Study

We conducted this study to examine the generalizability of our established effects on different group and screen sizes. Below, we detail each of these factors and then outline the study we conducted and our results.

*Group Size***.** In the studies reported thus far, we examined the effect of virtual vs. in-person interaction in the context of pairs. Supporting the substantive relevance of this context, survey evidence we collected indicates that the proportion of idea generation in pairs is sizeable. In a survey with a top multi-national management consulting firm, respondents report that 34% of their idea generation meetings were conducted one-on-one (see Supplementary Information section K). In addition to its prevalence, we selected the context of two-person groups because prior work finds that larger in-person group interactions (vs. pairs) can inhibit idea generation[36] and that larger virtual groups in particular struggle with conversation coordination (such as determining who should speak next)[37].

Nevertheless, whether our effect generalizes to larger groups remains an open and interesting question. In theory, many of the mechanisms underlying the reduced idea generation of larger groups (vs. pairs) demonstrated in prior research, such as social inhibition, production blocking, and social loafing [26,30,38], should also apply to videoconferencing. Further, prior research suggests that larger virtual teams uniquely struggle with conversation coordination because it is impossible to determine who the other team members are looking at[37]. Thus, we speculate that larger virtual groups would generate fewer ideas per person compared to virtual pairs and that both of these groups' performance would be dominated by the idea generation performance of in-person pairs.

However, to our knowledge, no prior research has empirically tested how group size affects idea generation performance in videoconferencing groups. Thus, to examine whether the established negative effect of 4-person (vs. 2-person) group size on idea generation productivity replicates in an online setting, we ran a videoconferencing study with two conditions: 2-person and 4-person groups (ideally, we would have conducted a 2 (in-person v. virtual) x 2 (2-person vs. 4-person) study, but this design was impossible due to the COVID-19 pandemic). Furthermore, to explore how group size affects virtual idea generation, we included measures corresponding to the mechanisms identified in prior in-person studies and examined how these constructs underly the productivity loss of larger groups online. Specifically, we focused on (1) fear of evaluation (such as social anxiety, heightened self-awareness, and fears of judgment), (2) production blocking (inability to voice ideas due to shared "airtime"), and (3) social loafing (reduced effort because each individual's performance is more dispensable in larger groups). We included other measures as well.

In addition to the mechanisms identified by prior research, we also predicted that another process, unique to larger *virtual* groups, would be at play. We expected that larger groups interacting online would have more trouble coordinating discussion because virtual groups do not have information about gaze between members of the group (one cannot observe who is looking at whom). To examine this novel mechanism, we included a measure of conversation coordination. In our pre-registration (linked below), we predicted that 4-person groups would experience more coordination issues than 2-person groups.

Screen Size. In our laboratory studies, we used 15.6" MacBook Pro Retina Display laptops because laptops are a common hardware used in videoconferencing, and 15.6" is the most prevalent laptop screen size offered in the market. In an online survey of people who currently use videoconferencing for work (N = 238), 81% of respondents reported using a laptop for video calls at least half the time. Further, of the participants who use a laptop for video calls, 74% used a laptop that was either the same size or smaller than the laptop size used in our studies (See Supplementary Information section N). These results indicate that the screen size in our studies is representative of people's typical experiences. Would our results generalize to larger screen sizes? Because we argue that virtual collaborators focus on the screen and filter out peripheral visual stimuli, the relevant comparison is the *ratio of screen size to room surface area*. If the screen takes up substantially more of the broader environment, virtual communication should require less visual focus, which in turn should attenuate our idea generation effect. However, even fairly large screen size options today do not take up a large amount of one's total environment. Thus, we suspected that the variance afforded by monitors and screens on the current market would not impact performance during idea generation. To test this, in this study, we measure participants' natural variance in screen size and examine if it relates idea generation performance in virtual pairs (even when controlling for income, comfort with videoconferencing, and time spent on the computer).

We pre-registered the design, sample size, and analysis plan for this study on AsPredicted (https://aspredicted.org/kb7xy.pdf).

Procedure

256 participants (76 male, 174 female, 4 non-binary, and 2 undisclosed, M_{age} = 24.7, SD_{age} = 8.19) from two university and staff pools participated in a video-conferencing study in exchange for $10. Of note, during data collection, we learned that some of the participants recruited using the student pool from a 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. Participants provided consent before beginning the study. This study was approved by the Stanford University Human Subjects Ethics Board (Protocol 35916).

We posted timeslots in an online research portal that allowed each participant to enroll anonymously into a group of four. The study was conducted by hypothesis-blind university research assistants who were not present during the group interaction. The design of this virtual study mimicked the study design of the first batch of data collection in the lab. Specifically, all participants joined a Zoom link (versions 5.0.0 and 5.0.1), and after confirming that each participant was on a computer and had both a functional video and audio, participants were informed their first task was to generate creative alternative uses for a Frisbee and that their second task was to select their most creative idea. As before, these tasks were incentive-aligned: each creative idea that was generated (as scored by outside judges) earned each participant one raffle ticket for a $100 raffle, and selecting a creative idea earned the participant ten additional raffle tickets. When assigning conditions, two thirds of the time, the four participants learned that they would be working together in one “video-conferencing room” as a 4-person group (N = 43), and one third of the time the participants were told that they would be working in two 2-person groups in separate video-conferencing rooms (N = 42). This allowed us to collect an equal number of 4-person and 2-person teams. After the group(s) entered a new “video-conferencing room” that did not include the experimenter, one participant from each group was randomly selected to be the typist (i.e., to record the ideas during the idea generation stage and indicate the selected idea in the idea selection stage for the pair) via a random number generator on Google.

In all conditions, participants were instructed to set up their screen such that half of the screen was a Google Sheet (accessed in 2021) with the instructions and the other half was the Zoom call. All participants were also instructed to hide their self-view (see Extended Fig. 5 for an example of what a participant's screen would look like in each condition). As in the lab study, each group generated ideas for five minutes and spent one minute selecting their most creative idea. Following the first batch of data collection in the lab, they indicated their top creative idea by putting an asterisk next to the idea on the Google sheet.

Once the groups completed both the idea generation and the selection task, each team member individually opened a survey and then logged off the videoconferencing call to complete the survey independently using Qualtrics (accessed in 2021). The survey measured screen size, conversation coordination, production blocking, fear of evaluation, social loafing, satisfaction with performance, along with other exploratory measures. Further the study included demographic measures, such as income, age, and comfort with using computers to control for endogeneity concerns regarding the screen size measure.

Dependent Measures

Idea Generation Performance. First, following the same procedure as the lab study, two undergraduate judges from the same population evaluated each idea on novelty and value. Judges again demonstrated satisfactory agreeability (α_{novelty} = .71, α_{value} = 60). Then, following our pre-registered plan, we divided the number of total ideas and creative ideas generated in each group by the group size to calculate the number of total ideas and creative ideas per person in each group. These measures determine which group size is more productive while controlling for number of participants who could potentially contribute ideas to the pool. This metric determines if it would be more productive to break up idea generators into groups of two or groups of four.

However, it is possible that two 2-person groups would generate a few redundant ideas, whereas an interactive 4-person group would not. Thus, in addition to our pre-registered plan, we also conducted a robustness check by creating randomized nominal groups of four (by combining two 2-person groups) and removing any repeated ideas. We then compared the number of ideas generated in the interactive groups of four compared to the nominal groups of four.

Idea Selection Performance. We followed the same procedure as the lab study to capture selection performance.

Results

Idea Generation. Because our main dependent variables are number of ideas and creative ideas per person, the dependent measures are not always integers, and thus, we cannot use a negative binomial regression to examine the effect of group size on idea generation. Instead, we examined whether our data meet the assumptions of a linear regression by evaluating whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distributions were normal by conducting the Shapiro-Wilk normality test[6]. The measures of total number of ideas and total number of creative ideas generated per person both violated the assumption of homogeneity of variance (Levene's test: *total ideas*: $F(1, 83) = 6.68$, $p = .011$; *creative ideas*: $F(1, 83) = 14.99$, $p <.001$) and were significantly non-normal (Shapiro-Wilk normality test, *total ideas*: $W = .94$, $p <.001$; *total ideas*: $W = .91$, $p <.001$). Thus, we analyzed the data using non-parametric tests.

Replicating prior research, 4-person groups generated significantly fewer ideas (M = 5.56, SD = 1.75) and significantly fewer creative ideas per person (M = 3.13, SD = 1.11) than 2-person groups ($M_{total\ ideas}$ = 9.50, SD = 3.20, $M_{creative\ ideas}$ = 5.38, SD = 2.21, *total ideas*: Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = 35.2, p < .001, Cohen's d = –1.53, 95% CI [–2.02,

–1.04]; *creative ideas:* Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = 26.2, *p* < .001; Cohen's *d* = –1.29, 95% CI [–1.77, –.82]). Of note, these effect sizes are similar to those reported in Mullen, Johnson, and Salas 1991's meta-analysis (*total ideas*: *d* = –1.40; *creative ideas*: *d* = –1.34).

We also examined the effect of group size on the average creativity of all ideas generated. First, we found that the variances of average creativity did not significantly differ by condition (Levene's test: *F*(1, 83) = .35, *p* = .557), but the distribution of average creativity was significantly non-normal (Shapiro-Wilk test: *W* = .96, *p* = .02). These results should be interpreted with caution because, with larger sample sizes, the Shapiro-Wilk test can produce a statistically significant results that are not practically significant[7]. Nevertheless, as a robustness test, we re-ran the analyses using a non-parametric test (which makes no assumptions regarding the underlying distribution).

We found that average creativity did not significantly differ by condition ($M_{2\text{-}person}$= 3.66, SD = .21, $M_{4\text{-}person}$ = 3.68, SD = .22, linear regression, N = 85 groups, *b* = .02, SE = .02, *t*(83) = .35, *p* = .731, Cohen's *d* = .08, 95% CI [–.36, .51]). Non-parametric tests provide the same conclusion (Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = .004, *p* = .951). Thus, there is no evidence that those in the 4-person condition came up with fewer but better ideas, compared to those interacting in pairs. See Extended Data Table 5 for summary.

As a robustness check, we also created nominal groups of four by randomly pairing groups of two together and removing any redundant ideas (N = 21). Using this alternative metric of group productivity, we again found that 4-person groups generated significantly fewer ideas (N = 43, *M* = 22.23, SD = 6.98) and significantly fewer creative ideas (*M* = 12.51, SD = 4.43) than nominal 4-person groups composed of two 2-person groups (N = 21, $M_{total\ ideas}$ = 30.67, SD

= 8.30, $M_{creative\ ideas}$ = 18.10, SD = 6.01, *total ideas*: negative binomial regression, N = 64 groups, *b* = –.16, SE = .03, *z* = –5.21, *p* < .001, Cohen's *d* = –1.15, 95% CI [–1.77, –.51]; *creative ideas:* negative binomial regression, N = 64 groups, *b* = –.18, SE = .04, *z* = –5.11, *p* < .001, Cohen's *d* = –1.14, 95% CI [–1.75, –.50]).

Idea Selection Performance. To ensure that our selection data meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. The difference score (Levene's test: *F*(1, 83) = .27, *p* = .603) and selected idea score (Levene's test: *F*(1, 83) = .69, *p* = .409) did not significantly violate the homogeneity of variance assumption, but the error score distribution was significantly non-normal (*W* = .95, *p* = .001). Thus, we will also include a non-parametric test result for the error score as a robustness check.

We find no evidence that group size influences idea selection. The scores of the idea that each group selected as their best did not significantly differ by group size ($M_{2\text{-}person}$= 3.88, SD = .96, $M_{4\text{-}person}$ = 4.09, SD = .79, Cohen's *d* = –.25, 95% CI [–.68, .19], *b* = .22, SE = .19, *t*(83) = 1.14, *p* = .256). Controlling for the quality of the top idea does not change the significance of this effect (*b* = .23, SE = .18, *t*(82) = 1.30, *p* = .196). In addition, the "decision error score" (difference between the creativity score of the selected idea and the creativity score of the actual top idea) was not significantly different by group size ($M_{2\text{-}person}$= 1.29, SD = .88, $M_{4\text{-}person}$ = 1.05, SD = .79, linear regression, N = 85 groups, *b* = –.24, SE = .18, *t*(83) = –1.32, *p* = .190, Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = 1.57, *p* = .211, Cohen's *d* = –.29, 95% CI [–.72, .15]). Lastly, the effect of group size on idea selection did not significantly change after controlling for number of ideas generated in each group (*decision error score*: linear regression, N = 85 groups, b = –.27, SE = .19, t(82) = 1.46, p = .148; *selected idea score*: linear regression, N = 85 groups, b = .26, SE = .18, t(80) = 1.39, p = .169).

Screen Size. The screen size of participants ranged from 9.5 inches to 34 inches. We averaged participants' screen size to calculate the average screen size of the group. The average screen size of a group was 15.11 inches, and the average group screen size ranged from 10.5 inches to 24.5 inches.

To examine the role of screen size on virtual idea generation, we ran regressions with the total number of ideas and total number of creative ideas generated per person as the dependent variables, screen size as the dependent variable, and group size as a covariate using a linear regression model. Of note, using group size as a covariate deviates from our preregistered plan. We opted to use it as a covariate because there was potential failure of random assignment and this could bias results due to the large effect of group size on number of ideas generated per person ($M_{2\text{-}person}$= 15.7, SD = 3.82, $M_{4\text{-}person}$ = 14.5, SD = 1.81, linear regression, N = 254 participants, b = –1.22, SE = .64, t(83) = –1.89, p = .062).

Screen size did not significantly correlate with number of ideas generated per person (linear regression, N = 85 groups, b = .15, SE = .09, t(82) = 1.59, p = .116; Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (66) = 71.93, p = .288) or number of creative ideas generated per person (linear regression, N = 85 groups, b = .05, SE = .06, t(82) = .82, p = .416; Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (66) = 70.68 p = .324). Given that screen size was not manipulated, which lends the analysis to endogeneity concerns, we ran pre-registered regressions with number of ideas and number of creative ideas generated per person as the dependent variable, screen size as the independent variable, and covariates of income and time spent on the computer (departing from our pre-registered plan, number of people in the room was not included as a covariate because, upon analysis, we realized it was often misinterpreted as a manipulation check (i.e., it was often interpreted as "how many people were in the Zoom room."; $M_{2\text{-}person}$= 2.72, SD = 1.43, $M_{4\text{-}person}$ = 4.20, SD = 1.30, linear mixed regression, N = 254 participants, *b* = .74, SE = .10, *t*(98) = 7.65, *p* < .001).

Importantly, even when including these covariates, screen size did not significantly correlate with number of ideas generated per person (linear regression, N = 85 groups, *b* = .13, SE = .10, *t*(78) = 1.34, *p* = .185) or number of creative ideas generated per person (linear regression, N = 85 groups, *b* = .05, SE = .07, *t*(78) = .68, *p* = .498). As a final robustness check, we ran regressions with number of ideas and number of creative ideas generated per person as the dependent variable, screen size as the independent variable, and covariates of income, time spent on the computer, comfort with Zoom, age, native language, and gender. Again, screen size did not significantly correlate with number of ideas generated per person (linear regression, N = 85 groups, *b* = .08, SE = .10, *t*(74) = .79, *p* = .431) or number of creative ideas generated per person (linear regression, N = 85 groups, *b* = .02, SE = .07, *t*(74) = .218, *p* = .828).

Finally, in an exploratory analysis, we examined if the effect of screen size on idea generation was moderated by group size. There was no significant interaction between screen size and group size on number of ideas generated per person (linear regression, N = 85 groups, interaction term: *b* = –.24, SE = .23, *t*(81) = –1.02 , *p* = .312) or number of creative ideas generated per person (linear regression, N = 85 groups, *b* = .09, SE = .17, *t*(81) = –.54, *p* = .588). Thus, we find no evidence that current screen size variation influences virtual idea generation.

Process Measures. To ensure that our measures meet the assumptions of a linear regression, we evaluated whether the data's variance was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. The Levene's test revealed that two categories violated the assumption of homogeneity of variance (*social loafing*: $F(1, 252) = 15.10$, $p < .001$ and *evaluation apprehension*: $F(1, 252) = 4.78$, $p = .030$). For these measures, we ran a non-parametric test. Further, the distributions of all measures were significantly non-normal (Shapiro-Wilk normality test, all Ws <.99, all ps < .05). Although this is not of great concern (as mentioned above), we will report results of both a linear regression model and a non-parametric test.

Replicating past research, we find that groups of four reported significantly more social loafing (reverse scored, as measured by an aggregated measure of ownership over ideas, perceived effort expended, and how personally responsible they felt to generate ideas, $\alpha = .74$, $M_{2\text{-}person} = 4.61$, SD = .73, $M_{4\text{-}person} = 4.25$, SD = 1.07, Kruskal-Wallis rank sum test, N = 85 groups, $\chi^2 (1) = 10.28$, $p = .001$, Cohen's $d = –.38$, 95% CI [–.65, –.11]) and evaluation apprehension ($\alpha = .84$, $M_{2\text{-}person} = 1.51$, SD = .78, $M_{4\text{-}person} = 1.67$, SD = .97, Kruskal-Wallis rank sum test, N = 85 groups, $\chi^2 (1) = 6.17$, $p = .013$, Cohen's $d = .18$, 95% CI [–.08, .45). Also, as we predicted in the pre-registration, groups of four reported more conversation coordination issues than groups of two ($\alpha = .72$, $M_{2\text{-}person} = 2.02$, SD = .87, $M_{4\text{-}person} = 2.30$, SD = .88, linear regression, N = 254 participants, $b = .28$, SE = .13, $t(100) = 2.26$, $p = .026$, Cohen's $d = .32$, 95% CI [.06, .59], Kruskal-Wallis rank sum test, N = 85 groups, $\chi^2 (1) = 7.41$, $p = .006$).

Finally, departing from prior research, we found that 2-person and 4-person virtual groups did not significantly differ in production blocking ($\alpha = .80$, $M_{2\text{-}person} = 5.95$, SD = 1.32, $M_{4\text{-}person} = 5.72$, SD = 1.50, linear regression, N = 254 participants, $b = –.24$, SE = .19, $t(252) = –1.22$, $p = .223$, Cohen's $d = –.16$, 95% CI [–.43, .10], Kruskal-Wallis rank sum test, N = 85 groups, $\chi^2 (1) = 2.85$, $p = .091$), or satisfaction with performance ($\alpha = .88$, $M_{2\text{-}person} = 5.16$, SD = 1.43, $M_{4\text{-}person} = 5.49$, SD = 1.14, linear regression, N = 254 participants, $b = .33$, SE = .19, $t(90)$

= 1.77, p = .080, Cohen's d = .27, 95% CI [0, .53], Kruskal-Wallis rank sum test, N = 85 groups, χ^2 (1) = .49, p = .482), although the directions are consistent with prior findings. We suspect that production blocking effect was not replicated due to more deliberate turn-taking that takes place in virtual interaction (supported by the fact that less crosstalk was observed in virtual interactions compared to in-person in the transcript of the lab study). We encourage future research to investigate this further.

Discussion

Overall, this virtual study demonstrated that the robust negative effect of group size on group idea generation replicates in a video context. The results of prior research, the results of this virtual lab study, and the results of our main studies all contribute to a broader understanding of group idea generation and modality. First, we know from a large body of prior work that pairs perform better at idea generation than larger groups *in person*[36]. Next, our virtual lab study confirms that pairs are better at idea generation than larger groups *virtually*. Lastly, our main studies show that in-person pairs outperform virtual pairs. These findings together culminate in our substantive recommendation: idea generation is optimized in pairs and in person

In addition, this study replicates established effects of group size on a variety of constructs (social loafing, illusion of productivity, and evaluation apprehension) in a video context. These results provide face validity to the self-report measures we used to rule out these alternative explanations in our main studies and provide further evidence that the modality of communication does not impact these important group processes.

N. Screen Size Survey

A total of 400 U.S. participants were recruited on a national online platform and completed a battery of items including an embedded screener question to determine eligibility: “Do you use video conferencing to meet with others for work?” 238 participants (143 men, 95 women; M_{age} = 36.6; SD$_{age}$ = 9.77) responded “yes” to this screener and were subsequently asked to participate in a follow-up survey for bonus payment. In this survey, we asked participants how often they used videoconferencing to meet with others for work (1 = *2-3 times a day*, 2 = *daily*, 3 = *2-3 times a week*, 4 = *once a week*, 5 = *once every 2-3 weeks*, 6 = *once a month*, 7 = *less than once a month*), what technology they use for communicating with others when videoconferencing (laptop, external monitor, or both), and if they used a laptop, what percentage of the time they used their laptop screen for videoconferencing. After, we asked participants to report the screen size of their laptop (ranging from 9 to 18 inches) and the screen size of their external monitor (ranging from 10 to 40 inches).

89% of participants used videoconferencing at least once a week, and 39% of participants used videoconferencing daily. During video calls, 47% of participants exclusively used their laptops, and 81% used their laptop at least half of the time. In contrast, only 7% used an external monitor exclusively. On average, people reported using their laptop screen for 70% of their calls. Importantly, we find no evidence that the type of screen (laptop vs. monitor) depended on whether or not the participant was working from home (88% of people who work from home use a laptop at least half the time, and 89% of people who do NOT work from home use a laptop at least half the time), although the vast majority of the sample worked from home at least part time (88% of participants) when this survey was conducted.

Of the people who use a laptop for videoconferencing, the mean laptop size was 14.5”. 74% of the people who use a laptop used a laptop that was either the same size or smaller than the one used in our lab studies (15.6”). These suggest that screen size in our studies is representative of people’s typical experiences. The average screen size of people who used a monitor for videoconferencing was 25.2 inches.

O. Heterogeneity Analysis

Prior work demonstrates that heterogeneous groups (e.g., groups that differ in background, experiences, or demographics) can more creative when they bring in "unique knowledge or association structures"[39]. For example, teams that share similar attitudes generate solutions that are significantly less creative than teams that possess differing attitudes[40]. To explore potential boundary conditions, we examined if the negative effect of virtual interaction on idea generation differed depending on the demographic heterogeneity of the groups.

Specifically, other than gender (which was explored below in Supplementary Information section P), we collected three different participant trait measures in each study that could be used to approximate trait heterogeneity in a pair: race, age, and whether they were a native speaker. Here, we measured the extent to which each pair was matched on these dimensions and examined whether that moderated the effect of condition on creative idea generation. We first calculated a "heterogeneity score" for both the race and native speaker items by assigning each pair a 0 if the pairs match (i.e., their race was the same) and a 1 if the pairs did not match (i.e., their races were different). For age, we calculated a heterogeneity score by subtracting the absolute value of the difference between the ages and dividing by the sum (e.g., if the ages are 25 and 30, the heterogeneity score would be (25-30)/(25+30) = .09). Thus, each heterogeneity score ranges from 0 to 1. We then averaged these three scores to generate a proxy for average heterogeneity in a pair (higher scores reflect more heterogeneity).

We first ran a negative binomial regression with number of creative ideas as the dependent variable, heterogeneity score (from 0 to 1) as the independent variable, and study as a covariate. The degree of heterogeneity in a pair did not significantly relate to creative performance (negative binomial regression, N = 300 pairs, b = –.08, SE = .12, z = –.66, p =

.507). We then examined if the effect of communication modality remains significant even when controlling for heterogeneity. Indeed, we find that communication modality significantly affects creative idea generation when controlling for the degree of heterogeneity in the group (negative binomial regression, N = 292 pairs, *b* = .16, SE = .05, *z* = 3.13, *p* =.002). In a subsequent exploratory analysis, we also examined if the effect of communication modality depended on heterogeneity. Specifically, we ran a regression with number of creative ideas as the dependent variable, heterogeneity score, communication modality condition, and their interaction as predictors, and study as a covariate. The heterogeneity score of the pair did not significantly moderate the effect of communication modality on creative idea performance (negative binomial regression, N = 292 pairs, *b* = .24, SE = .24, *z* = 1.01, *p* = .312).

The null effect of demographic diversity in our studies may appear surprising, but it is consistent with past work on diversity and creativity. Specifically, past work suggests that because heterogeneity benefits idea generation by providing more cognitive resources in the form of differing perspectives, demographic diversity should only impact idea generation when it is task relevant[41]. That is, demographic diversity facilitates group creativity only when different demographic groups provide differing perspectives on the task at hand. In our lab study, people of different races, ages, or native languages likely did not vary meaningfully in their knowledge structure of Frisbees or bubble wrap. Thus, more work is needed to explore the potential moderating role of task-relevant diversity in virtual groups.

P. Gender Composition Analysis

Prior work has demonstrated that women perform better in group tasks due to their interpersonal orientation[22,42,43]. For example, women often score higher on social sensitivity, cooperation, and turn-taking. Thus, as a test of generalizability, we examined if idea generation differed linearly depending on how many women were in the group (none, one, or two) or qualitatively between male-male, male-female, or female-female groups, and then tested if the effect of modality still emerges when controlling for gender composition in the lab study.

We ran a negative binomial regression with number of creative ideas as the dependent variable, number of female participants (0, 1, or 2) as the independent variable, and study as a covariate. Number of women did not significantly relate to creative performance (negative binomial regression, N = 292 pairs, b = .02, SE = .04, z = .58, p = .561). We then ran an ANOVA with creative ideas as the dependent variable and nominal gender composition (male-male, male-female, female-female) as the independent variable. In line with the prior regression, performance did not significantly differ by gender composition ($M_{f\text{-}f}$ = 7.35, SD = 3.12, $M_{m\text{-}f}$ = 7.31, SD = 3.82, $M_{m\text{-}m}$ = 7.20, SD = 3.02; F(2, 289) = .04, p = .965). Lastly, we examined if the effect of communication modality remains significant even when controlling for gender composition. Indeed, we find that communication modality significantly affects creative idea generation when controlling for number of females in the group (negative binomial regression, N = 292 pairs, b = .16, SE = .05, z = 3.11, p = .002) and when controlling for nominal gender composition (negative binomial regression, N = 292 pairs, b = .16, SE = .05, z = 3.11, p = .002).

In a subsequent exploratory analysis, we also examined if the effect of communication modality depended on gender composition. First, we ran a regression with number of creative ideas as the dependent variable, number of females, communication modality condition, and their interaction as predictors, and study as a covariate. The number of female participants in the pair did not significantly moderate the effect of communication modality on creative idea performance (negative binomial regression, N = 292 pairs, interaction effect: *b* = .07, SE = .06, *z* = 1.23, *p* =.220). As a robustness check, we also examined the interactive effect of nominal gender composition. Interestingly, we found that although the effect of communication modality is not significantly different between male-male ($M_{virtual}$ = 6.84, SD = 2.84, $M_{in\text{-}person}$ = 7.62, SD = 3.25) and male-female groups ($M_{virtual}$ = 6.17, SD = 3.32, $M_{in\text{-}person}$ = 8.27, SD = 3.97; negative binomial regression, N = 292 pairs, interaction effect: *b* = .17, SE = .16, *z* = 1.11, *p* =.267), the difference between virtual and in-person interaction for female-female groups ($M_{virtual}$ = 7.14, SD = 3.36, $M_{in\text{-}person}$ = 7.59, SD = 2.82) is significantly lower than for male-female groups (negative binomial regression, N = 292 pairs, interaction effect: *b* = .23, SE = .11, *z* = 2.04, *p* =.041). This suggests that perhaps the unique interpersonal orientation of female-only groups can mitigate the negative effects of virtual interaction. We encourage future research to explore this possibility.

Q. Other Industries

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 lab, generating alternative uses for a product, and a high-complexity task in the field, identifying both problems customers might have as well as generating solutions the firm could offer. Although it is reassuring that we demonstrate our effects across two different task types and participant pools who vary in their familiarity with each other, creativity training, and domain expertise, it is important to note that these populations are representative of only a subset of innovation teams. In particular, engineers have specific training in problem solving and idea generation[44] , attract certain personality types[45], skew male[45], and often engage in collaborative work[46]. Future research is needed to determine how our findings would generalize to other works contexts where employees might have different types of training or characteristics, such as in communications, education, fashion, healthcare, the military, and pharmaceuticals.

R. Average Creativity

We first evaluated whether the variance for our measure of average creativity was homogenous by conducting the Levene's test[5] and whether the distribution was normal by conducting the Shapiro-Wilk normality test[6]. The variance of average creativity did not significantly differ by condition in either study (lab study: $F(1, 299) = .55$, $p = .460$; field study: $F(1, 617) = .002$, $p = .964$).

Next, we only checked for normality in our lab studies because the central limit theorem states that, at sufficiently large sample sizes (N>100), the sampled means distribution will approximate a normal distribution even if the population significantly departs from a normal distribution[8]. Given that we had 619 pairs in our field study, our sample greatly surpasses the 100 sample-size threshold and thus it is unnecessary to check for normality. Our lab study also passes this threshold, however in the spirit of comprehensiveness, we report results from both a linear regression model and a non-parametric test (which makes no assumptions regarding the underlying distribution) for the lab study below, as the distribution is significantly non-normal (Shapiro-Wilk normality test: $W = .95$, $p <.001$).

Average creativity of ideas did not significantly differ by communication modality in the lab study ($M_{virtual} = 3.92$, SD = .41, $M_{in\text{-}person} = 3.95$, SD = .39, linear regression, N = 301, $b = .03$, SE = .05, t(299) = .60, $p = .547$; Kruskal-Wallis rank sum test, χ^2 (1) = .38, $p = .540$) or the field study ($M_{virtual} = 2.98$, SD = .40, $M_{in\text{-}person} = 2.99$, SD = .38, linear mixed effect regression, N = 619, $b = .01$, SE = .03, t(609) = .30, $p = .766$). An equivalence test with bounds of a negligible effect size (–.08 to .08) for the lab data revealed that the effect is statistically not different from zero ($t(299) = .64$, $p = .520$) and statistically not equivalent to zero ($t(299) = 1.07$, $p = .142$). An equivalence test with bounds of a negligible effect size (–.08 to .08) for the field data revealed that the effect is statistically not different from zero ($t(620) = 2.23$, $p = .013$). This suggests that the narrowed cognitive focus evoked by virtual communication reduces the generation of *all* ideas, good and bad[26,47,48].

S. Multiple Comparisons

No corrections were made for multiple comparisons in the main analyses or when examining alternative explanations. We did not make corrections for multiple comparisons for the main analyses because we only ran hypothesis-driven tests. We did not make corrections for multiple comparisons for the alternative explanation analyses because we wanted to conduct a more conservative test of whether other measures could explain our effect.

T. Supplementary Information References

1. Cameron, A. C. & Trivedi, P. K. *Microeconometrics : Methods and Applications*. (Cambridge university press, 2005).
2. Coxe, S., West, S. G. & Aiken, L. S. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment* **91**, 121–136 (2009).
3. Greene, W. Functional Form and Heterogeneity in Models for Count Data. in *Foundations and Trends in Econometrics* vol. 1 113–218 (Now Publishers Inc, 2007).
4. Vuong, Q. H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**, 307 (1989).
5. Levene, H. Robust tests for equality of variance. in *Contributions to probability and statistics* 278–292 (Stanford University Press, 1960).
6. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591 (1965).
7. Field, A., Miles, J. & Field, Z. *Discovering Statistics using R*. (Sage Publications, 2012).
8. Kirk, R. E. *Statistics: An Introduction*. (Thomson Wadsworth, 2008).
9. Gray, K. *et al.* “Forward Flow”: A New Measure to Quantify Free Thought and Predict Creativity. *American Psychologist* **74**, 539–55416 (2019).
10. Pennington, J., Socher, R. & Manning, C. Glove: Global Vectors for Word Representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014). doi:10.3115/v1/D14-1162.

11. M. A. West. The social psychology of innovation in groups. in *Innovation and creativity at work* 309–322 (John Wiley & Sons, 1990).

12. Nemiro, J. E. Connection in Creative Virtual Teams. *Journal of Behavioral and Applied Management* **2**, 93–115 (2016).

13. Ambady, N. & Rosenthal, R. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin* **111**, 256–274 (1992).

14. Cicchetti, D. V. & Sparrow, S. A. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency* **86**, 127–137 (1981).

15. Chartrand, T. L. & van Baaren, R. Human Mimicry. in *Advances in Experimental Social Psychology* vol. 41 219–274 (Elsevier, 2009).

16. Ireland, M. E. *et al.* Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science* **22**, 39–44 (2011).

17. Ashton-James, C. E. & Chartrand, T. L. Social cues for creativity: The impact of behavioral mimicry on convergent and divergent thinking. *Journal of Experimental Social Psychology* 5 (2009).

18. Gonzales, A. L., Hancock, J. T. & Pennebaker, J. W. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research* **37**, 3–19 (2010).

19. Berger, J. *et al.* Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing* **84**, 1–25 (2020).

20. Baltrusaitis, T., Mahmoud, M. & Robinson, P. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. in *2015 11th IEEE International*

Conference and Workshops on Automatic Face and Gesture Recognition (FG) 1–6 (IEEE, 2015). doi:10.1109/FG.2015.7284869.

21. Short, J., Williams, E. & Christie, B. *The Social Psychology of Telecommunications*. (John Wiley & Sons, 1976).
22. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **330**, 686–688 (2010).
23. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry* **42**, 241–251 (2001).
24. Gosling, S. D., Rentfrow, P. J. & Swann, W. B. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality* **37**, 504–528 (2003).
25. Bailey, E. R., Matz, S. C., Youyou, W. & Iyengar, S. S. Authentic self-expression on social media is associated with greater subjective well-being. *Nature Communications* **11**, 4889 (2020).
26. Diehl, M. & Stroebe, W. Productivity Loss In Brainstorming Groups: Toward the Solution of a Riddle. *Journal of Personality and Social Psychology* **3**, 497–509 (1987).
27. Paulus, P. B., Dzindolet, M. T., Poletes, G. & Camacho, L. M. Perception of Performance in Group Brainstorming: The Illusion of Group Productivity. *Pers Soc Psychol Bull* **19**, 78–89 (1993).
28. Hoffman, L. R. Applying Experimental Research on Group Problem Solving to Organizations. *The Journal of Applied Behavioral Science* **15**, 375–391 (1979).

29. Bradley, M. M. & Lang, P. J. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* **25**, 49–59 (1994).

30. Kerr, N. L. & Bruun, S. E. Dispensability of Member Effort and Group Motivation Losses: Free-Rider Effects. *Journal of Personality and Social Psychology* **44**, 78–94 (1983).

31. Kunda, Z. The case for motivated reasoning. *Psychological Bulletin* **108**, 480–498 (1990).

32. Zajonc, R. B. Social Facilitation. *Science* **149**, 269–274 (1965).

33. Bond, C. F. & Titus, L. J. Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin* **94**, 265–292 (1983).

34. Diehl, M. & Stroebe, W. Productivity Loss in Idea-Generating Groups: Tracking Down the Blocking Effect. *Journal of Personality and Social Psychology* **61**, 392–403 (1991).

35. Nemiro, J. E. The Creative Process in Virtual Teams. *Creativity Research Journal* **14**, 69–83 (2002).

36. Mullen, B., Johnson, C. & Salas, E. Productivity Loss in Brainstorming Groups: A Meta-Analytic Integration. *Basic and Applied Social Psychology* **12**, 3–23 (1991).

37. Vertegaal, R., Weevers, I., Sohn, C. & Cheung, C. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. *NEW HORIZONS* 8 (2003).

38. Camacho, L. M. & Paulus, P. B. The role of social anxiousness in group brainstorming. *Journal of Personality and Social Psychology* **68**, 1071–1080 (1995).

39. Paulus, P. Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology* **49**, 237–262 (2000).

40. Hoffman, L. R. Homogeneity of member personality and its effect on group problem-solving. *The Journal of Abnormal and Social Psychology* **58**, 27–32 (1959).
41. Bell, S. T., Villado, A. J., Lukasik, M. A., Belau, L. & Briggs, A. L. Getting Specific about Demographic Diversity Variable and Team Performance Relationships: A Meta-Analysis. *Journal of Management* **37**, 709–743 (2011).
42. Fenwick, G. D. & Neal, D. J. Effect of Gender Composition on Group Performance. *Gender, Work & Organization* **8**, 205–225 (2001).
43. Bear, J. B. & Woolley, A. W. The role of gender in team collaboration and performance. *Interdisciplinary Science Reviews* **36**, 146–153 (2011).
44. Kazerounian, K. & Foley, S. Barriers to Creativity in Engineering Education: A Study of Instructors and Students Perceptions. *Journal of Mechanical Design* **129**, 761–768 (2007).
45. Williamson, J. M., Lounsbury, J. W. & Han, L. D. Key personality traits of engineers for innovation and technology development. *Journal of Engineering and Technology Management* **30**, 157–168 (2013).
46. Salter, A. & Gann, D. Sources of ideas for innovation in engineering design. *Research Policy* **32**, 1309–1324 (2003).
47. Nijstad, B. A. & Stroebe, W. How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups. *Personality and Social Psychology Review* **10**, 186–213 (2006).
48. Keum, D. D. & See, K. E. The Influence of Hierarchy on Idea Generation and Selection in the Innovation Process. *Organization Science* **28**, 653–669 (2017).