Structuring Online Dyads: Explanations Improve Creativity, Chats Lead to Convergence

Faez Ahmed
University of Maryland
College Park, USA
faez00@umd.edu

Nischal Reddy Chandra
University of Maryland
College Park, USA
chandra.nischal@gmail.com

Mark Fuge
University of Maryland
College Park, USA
fuge@umd.edu

Steven Dow
University of California
San Diego, USA
spdow@ucsd.edu

ABSTRACT

Exposing people to concepts created by others can inspire novel combinations of concepts, or conversely, lead people to simply emulate others. But how does the type of exposure affect creative outcomes in online collaboration where dyads interact for short tasks? In this paper, we study the creative outcomes of dyads working together online on a slogan writing task under different types of interactions: providing both the partner's idea and their explanation for that idea, enabling synchronous chat, and only exposing a person to their partner's idea without any explanation. We measure the creative outcome and define text-similarity-based metrics (e.g., mimicry, convergence, and fixation) to disentangle the interactions. The results show that having partners explain their ideas leads to largest improvement in creative outcome. In contrast, participants who chatted were more likely to reach convergence on their final slogans. Our work sheds lights on how different online interactions may create trade-offs in creative collaborations.

CCS Concepts

•Human-centered computing → Computer supported cooperative work; Empirical studies in collaborative and social computing;

Author Keywords

Creativity; distributed teams; chat; design examples; creative collaboration; computer-mediated communication

INTRODUCTION

For many problems, people work together in teams to generate creative solutions. Many empirical studies on team performance have been conducted to examine individual attributes, such as personality types, social/emotional intelligence, ability level, and cultural values in teamwork [26, 20, 23, 42]. Researchers have also studied how teams interact, both when collocated or working remotely, and provided guidance on the choice of tools for different collaborative tasks [9]. A key question in team performance is, how should teammates interact online to best improve creative outcomes?

In this paper, we focus on interaction mechanisms that affect how dyads collaborate on online tasks to generate creative ideas. We compare different types of direct and indirect interaction between pairs of online participants. Specifically, we explore how dyads perform tasks with increasing level of interaction under following four conditions: 1) Expose: When team members interact indirectly by viewing each other's idea but do not communicate. 2) Explain: When team members view each other's idea but also give a written explanation of their idea to their partner. 3) Chat: When team members view each other's idea and then ideate together by chatting with their partner about their ideas. 4) Discuss: When team members view each other's idea and explanation for the idea and then ideate together by chatting with their partner.

Our analysis compares these different types of interaction on group-level creative outcomes. We use text similarity-based metrics to uncover how users interact and behave in different conditions, including whether dyads converge or diverge from each other. Our findings show that: 1) Explaining one's idea inspires teammates to generate more creative ideas than the conditions involving synchronous chat; 2) teammates are more likely to converge on similar ideas when they chat; and 3) text similarity measures can identify unproductive chats.

This paper makes the following main contributions:

1. We study how the types of interactions between dyads affect creative output. Specifically, we show that idea-explanations lead to more creative final outcomes, but that chat promotes convergence among teammates.
2. We analyze chat logs to unpack how mutually productive chats compare with chats dominated by one person's idea.
3. We discuss implications for supporting computer-mediated creative work and improving interaction around examples through chat and idea-explanation.

C&C '19, 2019 Creativity and Cognition, San Diego, CA, USA June 23–26, 2019

DOI: 10.1145/3325480.3325486

RELATED WORK

We review existing research on brainstorming, including the pros and cons of exposing others' ideas. Next, we discuss efforts to study interaction between participants, specifically synchronous chat and idea explanations.

Brainstorming in teams

Generating creative ideas is important for any enterprise to innovate. Prior work on brainstorming shows the importance of asking people to independently create ideas before exposing them to others. Although brainstorming is expected to enhance the number and quality of the ideas generated, controlled studies that compare interactive brainstorming with nominal groups have shown that verbal brainstorming in groups actually hinders the number of ideas generated [38, 35, 30]. This has been attributed to a range of factors such as reduced motivation in groups (i.e., social loafing), concern with evaluation of ideas in groups, group members matching their performance to that of the low performers in the group or the fact that only one idea can be expressed at one time in the group [45].

Electronic brainstorming has been proposed to counter some of these problems [15]. Past work on group electronic brainstorming has been reported to provide significantly more original ideas than nominal groups. Baruah *et al.* [6] report that electronic brainstorming groups became more creative and exhibited slower productivity loss compared with the nominal groups working electronically over time. Paulus *et al.* [45] show that electronic group interaction enhances the number of ideas and their quality relative to similar number of individuals who generate ideas without sharing them. The authors also show how increasing the number of examples shown to individuals affect these outcomes. Interactive groups produce fewer combinations because viewing and thinking about others ideas takes time away from production [31]. Our research adopts the nominal group approach for generating initial ideas, but then we pair people up to share ideas in order to compare different models for interactive brainstorming.

Pros and cons of exposing people to ideas

Exposing ideas by others often affects one's ability to ideate [34, 4]. Such exposure may have motivational effects since individuals may use the ideas generated by others as a reference point for their own performance (social comparison effect). For example, Nojstad *et al.* [40] showed that individuals exposed to others' ideas performed better than those who were not. Exposure to others' ideas can provide a cognitive spark, for example in work by Chan *et al.* [12], experts monitor incoming ideas from the crowd through a dashboard, and curate high-level "inspirations" to guide ideation towards interesting solution themes. Similarly, Andolina *et al.*[3] proposed the Crowdboard system, where in-person ideators can elicit synchronous creative input from online crowd workers recruited on-the-fly. In contrast, our work focuses on comparing different modes of interacting and exposing ideas between two people working together online.

The disadvantages of showing examples can be that reading ideas of others can limit the time one has to generate one's own ideas [28]. Exposure to ideas from others may also increase idea uniformity since these ideas may prime similar ideas [45]. Exposure to a large number of ideas may result in cognitive overload and a tendency to ignore the presented ideas. Researchers also found that groups tend to focus on agreement instead of diversity of perspectives [49].

When exposed to other ideas, individuals may exhibit fixation, which is considered an impediment to productive problem solving [32]. In interactive brainstorming, rather than exploring a diverse set of ideas, participants might conform to the categories of ideas suggested by other group members. Empirical studies in design have also shown exposure can cause fixation [17, 56].

Many methods have been proposed to decrease fixation during ideation exercises [28, 47]. To help prevent people from prematurely conforming to existing ideas, some research would recommend that groups alternate between generating ideas alone and being exposed to their peers' ideas [46]. This paper explores interactions that enable people to first work independently and then effectively communicate with others such that it reduces fixation and improves creative outcomes.

Improving collaboration with synchronous team chat

Synchronous chat allows individuals to not only explain their own ideas to online collaborators, but also discuss new ideas and to potentially riff on each other in real time. While chat might help some teams creatively combine their ideas, others might feel prohibited in their interaction and only shallowly influence each other [16]. Social loafing [30] may also occur in synchronous chat if individuals give less effort to the team task due to diffused responsibility.

To analyze how chat affects outcomes, Coetzee *et al.* studied online, synchronous, interactive peer learning for both crowdworkers and students [14]. They found that teams who chatted not only answered more correctly than individuals but also enjoyed the experience more. Furthermore, participants who justified their answers improved further. Similarly, Niculae *et al.* explored whether interactions can help improve workers' output by proposing a framework for analyzing conversational dynamics to determine whether a given task-oriented discussion provides value [39]. They focus on applying natural language processing techniques to predict whether a discussion would be constructive based on analyzing users' chat logs. They tag a team discussion as 'constructive' if it results in an improvement over the potential of the individuals. By doing so, the authors show that factors like balance in idea contributions between the team members is a good indicator of productive discussions. We use a similar measure of balance to identify non-productive chats. While studies in collaborative work [53] and sensemaking methods [22] have studied information sharing for specific interaction conditions (say chat), our study compared how different interaction conditions affect the creative outcome and is the first to explore the cross-product between chatting and explaining.

Many studies on conversational processes have studied the linguistic makeup of conversation logs and used dictionary text features (like LIWC) to provide feedback during conversation

[50] or to predict performance [21]. In contrast, our measures for similarity calculation (universal encoders) do not assume pre-defined dictionary features. We use a state-of-the-art word embedding method for all metrics, which improves upon on past metrics that use a TF-IDF representation scheme [52, 53].

Improving collaboration with idea explanations

To expose the underlying thinking behind an idea, team members could be encouraged to explain their idea, which could be done asynchronously, or used in conjunction with real-time chat. Prior work shows that explaining an idea often helps the recipient incorporate the new information with their own [41].

Researchers have also explored the idea of peers explaining their idea to partners. For instance, Drapeau proposed a MicroTalk workflow for crowd teams *et al.* [19]: given a labeling task, workers first assess the task and give their answers independently; workers are then asked to come up with arguments to justify their answers; finally, workers are presented with arguments from a different answer and are then asked to reconsider their answers. Similarly, Chang *et al.* [13] propose a three stage procedure to improve label accuracy. This body of results imply that including participant explanations could have positive impacts on team outcome. However, prior work has focused on relatively simple tasks and the worker interactions are limited to presenting arguments from another worker. In contrast, our work compares different modes of interaction on a creative task.

Inspired by the effects that different types of interactions may have, we examine the relative effects of enabling chat and encouraging explanations, or using both, on the creative outcomes of dyads. To do so, one option is to independently create and then share ideas with others. Dow *et al.* [18] show that dyads who share multiple ideas, as opposed to just one idea, more effectively explore the idea space and produce more creative outcomes. A second option, and one used in recent studies [14, 13], is to ask participants to reflect on their own idea and provide explanation or justification for it. Finally, a third option, inspired by electronic brainstorming [46], is for team members to discuss their ideas through synchronous chat. In our study, we compare and contrast these options and explore the underlying factors affecting dyad interactions.

METHOD

Experimental design

We designed a 2×2 study with four interaction conditions: ‘Expose’, ‘Explain’, ‘Chat’ and ‘Discuss’ as shown in Fig. 2. The variation between the conditions occur on two axes—whether participants are allowed to chat or not and whether their idea-explanation reaches the partners or not. The four conditions allow us to understand how emphasizing these elements of interaction affects creative output.

After completing the ideation task alone, each participant encounters one of the four interaction conditions for the team task. Participants do not know their condition apriori or if they will get to chat with a partner or not. The four conditions are:

- Expose: Partners are not allowed to chat. After the selection phase, each participant is shown her favorite idea along with her partner’s favorite idea.
- Explain: Partners are not allowed to chat. After selection, each participant is shown her favorite idea along with her partner’s favorite idea. Compared to Expose condition, each participant also views the explanation her partner provided.
- Chat: Partners are allowed to chat with each other. Similar to the Expose condition, each partner is shown her selected idea along with her partner’s selected idea. A chat window opens within which she can live chat with her partner to brainstorm or refine their ideas.
- Discuss: Partners are allowed to chat with each other. Similar to Explain condition, each partner is shown her favorite idea along with her partner’s favorite idea. A chat window opens within which the partners can chat synchronously to brainstorm or refine their ideas.

The precise prompt shown to participants in the Chat condition is: “Read your partner’s slogan. Now chat with your partner and write a slogan for the product before the timer runs out. Create new ideas that blend the strengths of your idea and your partners, don’t just copy/paste one of them.”

Participants

We recruited 120 participants from Amazon Mechanical Turk where each condition received thirty unique participants paired randomly to work in teams of two. Each worker who accepted the HIT was paid $0.35 as standard payment to stay in the waiting room before their partner arrives. Workers were asked to continue with other tasks and keep the window open. As soon as another worker accepted the HIT, those two were paired into a team for the main task. We give a $2 bonus to each team member for completion of the main task. The average completion time for all tasks is 8.3 minutes, leading to an hourly wage of $17. Introducing direct communication between pairs of participants on the same tasks required us to synchronize the work pace of pairs of participants [7, 8]. To overcome the challenge of pairing up crowd workers to perform real-time collaborative work, we adopted a “waiting room” strategy [33]. Workers who dropped out mid-task were given bonus in proportion of time spent and phases completed. The average age of participants in our experiment was 37 years, 56% of which were males and most common ethnicity was white (78%). Bachelor’s degree was the most common highest level of education among workers. 62% workers reported being employed full time.

Procedure

We conducted a randomized experiment on MTurk to test our hypotheses regarding how interactions affect a team’s creative outcomes. Participants were paired into teams of two to work on a creative ideation task. Figure 1 shows the workflow of a typical task given to a participant. Each task includes seven phases—1) Matching, 2) Pre-survey, 3) Ideation, 4) Selection, 5) Interaction 6) Final and 7) Post-survey.

In the matching phase, participants who accept the HIT are randomly matched to form two-person teams. Once the team

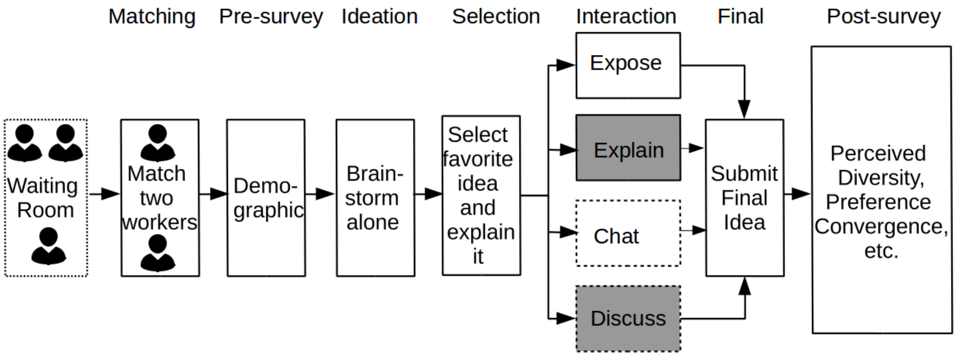

Figure 1. Task workflow for MTurk experiment to compare four conditions. In the interaction phase, greyed box shows conditions where explanation of idea is displayed to partner. Boxes with dashed outline denote conditions where partners can participate in live chat.

	Cannot See Explanation	Can See Explanation
No Chat	Expose	Explain
Allows Chat	Chat	Discuss

Figure 2. The four conditions considered in our study. Teams in the conditions in the right column see an explanation from their partner, while the teams in the bottom row interact via chat.

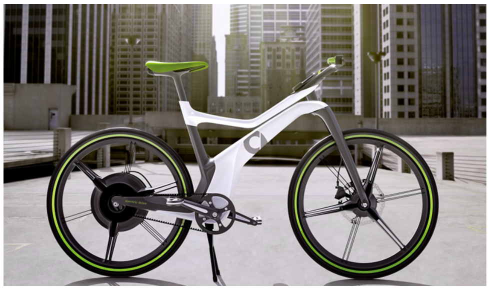

Figure 3. Ideation task: Write a three sentence advertising slogan for this transportation device. This slogan will be shared with your partner for the team task. You will spend 4 minutes on this task until the countdown runs out.

is formed, each partner is notified that they have been matched and the task starts. In the pre-survey phase, each participant independently completes a survey containing questions about a participant's demographics. Once the participant completes pre-task surveys, she is directed to the ideation phase where she works alone on an ideation task: generate advertising slogans for a futuristic looking bike. The participants can enter as many ideas as they want and we ask them to work on this task for a minimum of four minutes. After entering all ideas, each participant moves to the selection phase where she is shown all her submitted ideas and asked to select her favorite idea. She is informed that this favorite idea will be shown to her partner and asked to provide explanation or justification supporting her idea.

We designed the experiment such that all participants are asked to explain their favorite idea regardless of condition. However, this explanation is only displayed to the partner in two of the four conditions in order to tease apart the effects of writing explanations (which all participants did) and showing the explanations. At the end of the interaction phase, each partner independently enters a final idea in the final phase—*e.g.*, this can be a repeat or modification of a previous slogan, or a completely new slogan. After submission, they proceed to the post-survey phase, which has survey questions to measure their perceived diversity, helpfulness and enjoyment scores as discussed in following sections.

Ideation Task Design

For ideation phase, participants generated advertising slogans for a transportation device shown in Fig. 3. We chose this ad design task because it fulfilled several key criteria discussed in [37, 18]: 1. Participants could exhibit individual creativity, but it would also benefit from collaboration. 2. It is open-ended, complex and accepts different viewpoints, thus it is likely to be affected by interpersonal dynamics. 3. It could be completed in a short duration and did not require previous knowledge. 4. External judges could rate the quality of the work.

Participants could enter as many slogans as they wanted and had to spend minimum four minutes on the ideation task with no maximum time limit.

Measures

We capture multiple attributes from participants, either by directly asking them a survey question or by analyzing their text input. We explain these measures below.

Creative outcomes: We measure creativity of a slogan from expert ratings on novelty and quality. For any given idea, we ask experts to rate on a five point scale (Not at all to Very much) for two questions: 1) How unique, unusual, or novel is this slogan? 2) How useful is this slogan for the intended purpose?

These survey questions are based on prior work [44], where novelty and quality questions were used to find concepts which are more creative. For novelty, we also clarified to the reviewers that they should focus on global novelty and not local novelty (novelty compared to other slogans within a particular survey). We take the average of novelty and quality ratings to calculate the creativity score by mapping the Likert scale

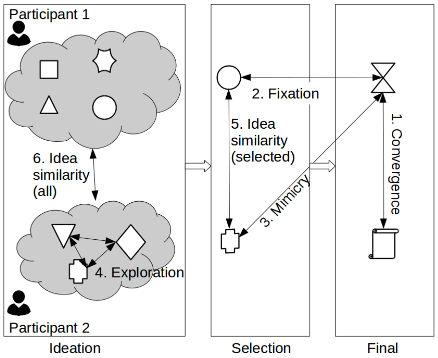

Figure 4. Text similarity metrics calculated for ideas. Mimicry is defined as the similarity of Person 1's final idea and Person 2's selected idea. Convergence is the similarity of two final ideas, while fixation is similarity between same person's final idea and selected idea.

responses to equally spaced intervals between 0 and 4, with 4 being the highest possible creativity. By taking average ratings from multiple raters, the score is expected to provide a signal of true creativity of the slogan. We measure creativity of the slogan selected as favorite in selection phase (named $Creativity_s$) and also the final slogan from the final phase (named $Creativity_f$). To measure improvement of an individual's slogan, we calculate the difference in creativity scores, termed as Δ creativity, which equals $Creativity_f - Creativity_s$.

Perceived Diversity: Creative output of a team can be affected by diversity of the team. However, concept of "diversity" has a variety of meanings, including separation in attitudes or viewpoints; variety of positions, categories or backgrounds; and disparity in values on some resource or asset [24]. We measured perceived diversity after completion of the interaction task to understand how different users perceived their partners from themselves. Six items measuring perceived diversity were taken from previous research [25]. Sample items include "My partner and I were similar in priorities" and "My partner and I were similar in commitment to working hard on this task." Similar to [55], we reverse-coded the score on the five point scale to get the measure of perceived diversity. We asked these questions in the post survey phase after completion of the final phase. Factor analysis was performed, and showed that the questions used loaded on one and the same factor and Cronbach's alpha was 0.89, which indicates a reliable reflective variable.

Helpfulness, Enjoyment and Motivation: In addition to above measures, we ask six additional questions in the post survey to judge how the users felt about the task and their experience with partner interaction. To measure helpfulness of partner, we ask users to answer on a scale of 1 (Disagree strongly) to 7 (Agree strongly)—a) "My partner's comments were helpful" and b) "My ideas and comments were helpful". To measure task enjoyment, we ask users to answer on a scale of 1 (Disagree strongly) to 7 (Agree strongly)—a) "I enjoyed working with my partner on this task" and b) "My partner and I worked well together." To measure their will to perform better on task, we ask users to answer on a scale of 1 (Disagree strongly) to 7 (Agree strongly)—a) "I wanted to perform well on this task" and b) "My partner wanted to perform well on this task". As discussed in the results section, these factors indicate how satisfied a participant is with the team task.

Similarity measures based on text: Figure 4 shows different text based metrics derived from responses given by participants. The top part shows the ideas submitted by the first user in different phases and the bottom part shows the ideas submitted by her partner. To measure relationships between these ideas, we use Universal Sentence Encoder [11]—one of the state-of-the art text embedding methods to find vector representation of each slogan. The Universal Sentence Encoder encodes text into high dimensional vectors and is widely used for text classification, semantic similarity, clustering and other natural language tasks. The models are trained on varied sources like Wikipedia, web news, web question-answer pages and discussion forums and outputs a 512 dimensional vector for each slogan.

Similar to Semantic Textual Similarity shared tasks [10], we used cosine similarity to estimate the relatedness of each pair of text ideas. The similarity score is 1 if two text responses are exactly the same. To verify these ratings, we asked four human raters to score similarity between a set of 45 slogans (generated in our trial experiments) on a scale of 0 to 5. We calculated similarity scores given by word embedding method and calculated the correlation of these scores with the human-ratings using Pearson's correlation coefficients. Similar approach has been used in literature [51] to verify sentence embedding methods, where correlation values between 0.4 to 0.7 was observed across different automated methods. We found our method's correlation with average human similarity ratings to be 0.71, showing that it captured similarity between text slogans.

We use text similarity values to calculate the following measures: Convergence, Fixation, Mimicry, Exploration, Idea similarity (selected) and Idea similarity (all) as shown in Fig. 4, and describe how they may help in uncovering interactions between partners.

1. **Convergence:** The convergence metric denotes how close the final ideas of the two partners were. We measure convergence as the similarity between the submitted 'Final Idea' for both participants. If the convergence score is closer to one, it indicates that both partners submitted the same final idea. We hypothesize that with increasing interaction, convergence of ideas should increase as partners can gain consensus on a final idea.

2. **Fixation:** In creative problem solving, individuals often face fixation, an impediment to productive problem solving [36]. To measure fixation, we calculate the text similarity between a user's favorite idea (pre-treatment) and her final submitted response (post-treatment). If the similarity is closer to one, it indicates a possibility of fixation as the person did not change her initial response after the interaction activity.

3. **Mimicry:** We define mimicry as a measure of someone imitating their partner's idea. Aston *et al.* [5] found that mimicry facilitates creative problem solving by increasing convergent thinking, which aids the identification of a sin-

gle, common solution from multiple alternatives. However, it impedes generation of novel ideas. As shown in Fig. 4, we measure mimicry as the similarity between participant's final idea and her partner's favorite idea shown to them. A mimicry score close to one indicates a possibility that the person simply imitated her partner's favorite idea shown to her. Note that our measure named 'mimicry' is a somewhat different use of the word than in the work of Scissors *et al.*[48] and Gonzales *et al.*[21], who study convergence on linguistic dimensions (such as vocabulary) as a way of signaling affinity toward partners.

4. **Exploration:** Exploration is defined as how well the idea space is covered. We measure exploration score for a user by measuring the average pairwise dissimilarity between all ideas generated by a user in the ideation phase. If a user generates multiple ideas similar to each other, then her score will be close to zero, while if she generates multiple ideas which are quite different from each other, the set will get a higher exploration score. Users who generate only one idea get a score of zero. As the number of ideas increase, we expect the exploration score to increase. For the same number of ideas, higher exploration score will indicate a more diverse set of ideas submitted by the user.
5. **Idea similarity (selected)**: To measure the distance between ideas of partners, we calculate the cosine similarity between their favorite ideas. As individuals view each other's favorite idea, similarity between user's and her partner's idea can affect openness to adoption of new idea. On the other hand, quite dissimilar ideas can open new directions of thought for the user.
6. **Idea similarity (all)**: Although individuals view each other's favorite idea, they generate multiple ideas in the ideation phase. The similarity between all ideas of one participant to all ideas of another participant may indicate how similarly they thought about the problem overall (irrespective of what slogan they chose as favorite). To measure it, we calculate the cosine similarity between average vector representation of all ideas of two partners.

Word count: Past literature has indicated that the length of a text artifact (total number of words) often correlates with quality ratings [2]. While we do not expect word count to correlate with idea quality for our experiments, we measure it to understand differences across interaction conditions. We define wc_f as the total number of words in idea submitted in final phase and wc_s as the total number of words in the favorite idea of a user. Δwc measures the increase in the number of words from selection to final stage.

RESULTS

Participants in our experiments submitted a wide variety of slogans. The explanations provided for slogans often provided details on what the slogan meant. For instance, one participant in Chat submitted the slogan:

> The best, the brightest and the most beautiful, class on two wheels.

They then explained:

> I think this bike would be purchased often by Millennials. They don't care about price, they care about how something looks, who it impresses and how well it functions. I think this is a beautiful bike with excellent workmanship (the best) and very classy. It is made to impress and to function. Considering who we are marketing it to, I think stressing those points and yet keeping it simple is the way to go.

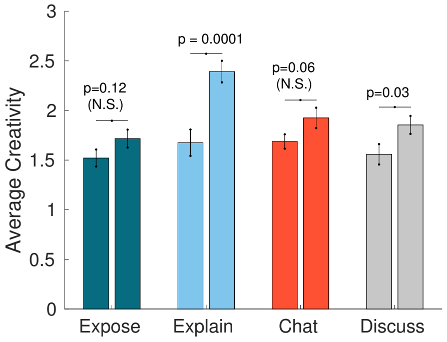

Figure 5. Participants in the Explain and Discuss conditions saw a significant shift in average creativity after interaction with the partner. For each condition, left side bar indicates the creativity of the favorite slogan from the selection phase and the right side shows the average creativity of the final slogan.

Final creativity is higher in the Explain condition

To calculate the creativity for each condition, we first find the average creativity of each slogan from four raters. Next, we find the average ratings of all thirty slogans in a condition to calculate the mean creativity for each condition. Fig. 5 shows the average creativity ratings for the four conditions. Bar on the left within each condition represents the average creativity of slogan generated before intervention, while the one on right indicates the average creativity of final slogan. We find that average final creativity is highest for the Explain condition (mean (μ) 2.39, standard deviation (SD) 0.59), followed by the Chat (μ 1.85, SD 0.49), the Discuss (μ 1.92, SD 0.55) and the Expose (μ 1.72, SD 0.49) conditions.

To statistically compare the final creativity in the four conditions, we performed a four-way comparison between Expose, Explain, Chat and Discuss using analysis of variances. ANOVA found a significant main effect between the four cases for final creativity (p value 2.36e-05). We followed up with Tukey's test (with a family-wise error rate of 0.05) to investigate the differences in means between the three conditions. The differences in means are shown in Table 1, which shows that the final creativity is highest for Explain and least for Expose. A post-hoc Tukey HSD test showed that the pairwise differences between Explain and other three conditions is significant with mean difference of 0.68 for Expose, mean difference of 0.54 for Chat and 0.47 for Discuss.

Since we define creativity as the average of quality and novelty, we also compare the attributes individually and find that Explain achieves both higher quality and novelty. An ANOVA followed by Tukey's test showed that the results are statistically significant for all comparisons for both novelty and quality (except for comparison of average quality between Explain and Chat, although the mean difference is large). As quality, novelty and creativity are correlated (Pearson correlation > 0.9), we only report the creativity results in Table 1.

Factor	ANOVA F value	ANOVA p value	Mean Difference Discuss - Chat	Mean Difference Explain - Chat	Mean Difference Expose - Chat	Mean Difference Explain - Discuss	Mean Difference Expose - Discuss	Mean Difference Expose - Explain
Creativity$_f$	8.89	2.36e-5	0.07	-0.47*	0.21	-0.54*	0.14	0.68*
Δ creativity	3.93	0.01	-0.05	-0.48*	0.04	-0.42	0.10	0.52*
Δ word count	3.36	0.02	-4.7*	-12.67	-1.63	-7.97	3.07	11.03
Perceived diversity	2.78	0.04	-0.51	-0.65*	-0.46	-0.14	0.05	0.19
Convergence	14.90	2.86e-8	-0.02	0.24*	0.24*	0.26*	0.26*	0.00
Helpfulness	2.80	0.04	0.83*	0.58	0.65	-0.25	-0.18	0.07

Table 1. Attributes where significant differences were found. A score of -0.47 between Explain and Chat for *Creativity$_f$* implies that mean final creativity of Explain is 0.47 higher than Chat. Significant difference in mean from Tukey's test are indicated by *.

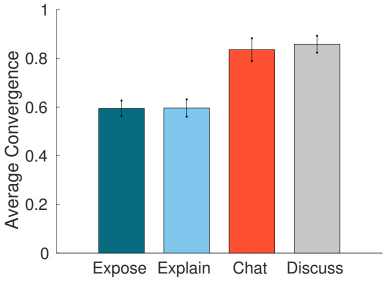

Figure 6. The average convergence is higher when the teammates had an opportunity to interact in real-time.

Creativity increased more in the Explain condition

The interaction conditions not only affects the final creativity but also the change in creativity. Hence, we compared the Δ creativity for all the four conditions. The Δ creativity is also highest for Explain condition (μ 0.72, SD 0.76). Tukey's test indicates significant pairwise differences in Δ creativity for both Explain vs Discuss (mean difference 0.42) and Explain vs Expose (mean difference 0.52).

Real-time interaction helped dyads converge

Figure 6 compares the final slogan convergence across the four conditions. We expected and observed that chatting allows members to come to agreement on a common slogan, leading to significantly higher convergence scores for Chat and Discuss compared to Expose and Explain.

We also notice the effect of higher interaction in our post-survey questions, where ANOVA found a significant difference for perceived diversity and helpfulness. The perceived diversity (μ 0.80, SD 0.91) is lowest in the Chat condition and users in this condition also report highest average helpfulness score (μ 6.1, SD 0.99). The difference in mean shows that users in the Chat condition found their partner more helpful and perceived them as more similar to themselves compared to Expose, Explain and Discuss. However, the difference in perceived diversity is only significant between Explain and Chat. The difference in helpfulness is only significant between Discuss and Chat. Surprisingly, despite similar chat based interactions, the Discuss condition did not have similar perceived diversity and helpfulness scores to that of Chat. There are two possible explanations for this difference. First, as members have more information to process in Discuss condition, they may not have time for interactions. This would lead to fewer chats and lesser team satisfaction with the chat activity. Second, it is possible that the higher information provided in Discuss (via self explanations) lessened the need for members to interact further for the team task. Hence, they did not need to communicate as they already had enough information to complete the task. By analyzing the chat logs, we show later that there is more evidence supporting this reason.

Differences between conditions for other factors

Other differences across conditions were not significant. Although the average mimicry for Discuss (μ 0.63, SD 0.22) and Chat (μ 0.61, SD 0.20) was higher compared to Expose (μ 0.59, SD 0.21) and Explain (μ 0.55, SD 0.22), no statistical difference is found between the four conditions for mimicry scores. Similarly, the average fixation for Discuss (μ 0.72, SD 0.19) and Chat (μ 0.70, SD 0.16) is higher compared to Expose (μ 0.65, SD 0.20) and Explain (μ 0.62, SD 0.20) but the differences are not significant. We did not find statistical differences between conditions for average initial creativity of favorite ideas, average total number of ideas generated, exploration score and partner similarity too.

Note that in comparing the different factors in Table 1 using ANOVA, we assume that partners are independent, as we instructed them to submit their slogans independently. However, it is possible that the two slogans may depend on each other by a factor which is hard to quantify. To account for this possible issue in statistical comparison, we conducted a followup test by randomly sampling 15 individuals from each condition, such that no two partners are selected in a sample set. We collect 1000 such samples. Next, we do a one-way ANOVA to compare the conditions for each of the 1000 sample set. Where the null hypothesis is rejected, we conduct a followup Tukey's test. On running this simulation, we find that the conditions differ significantly ($p<0.05$) for Convergence (99.9%) and Final creativity (93.3%) in a large proportion of the comparisons. However, other factors like perceived diversity (42.1%), increase in creativity (29.1%) and increase in word count (32.3%) are not significant in majority of the random samples. However, by sampling half the members, we reduce the statistical power of the test too.

Correlation analysis

We calculate Pearson's correlation between all measures combined. There was positive correlations between novelty final and quality final (correlation 0.68) as well as novelty selected and quality selected (correlation 0.57). This indicated that slogans which received higher quality ratings received higher novelty ratings too. We also found perceived diversity was negatively correlated with enjoyment ratings (-0.82). On the other hand, enjoyment ratings and helpfulness were highly correlated (0.81). This showed that in all cases, users who perceived their partners as different from themselves reported enjoying the task less and found their partner to be less helpful.

When we focus specifically on Δ creativity for each condition separately, we find that four factors (Δ word count, enjoyment, helpfulness and motivation to perform well) have the highest correlation with an increase in creativity score among all conditions. The correlations for last three of these factors indicates that an increase in creativity was accompanied by participants who found their partner's input helpful, enjoyed working with their partners, and felt that the team wanted to perform well (the correlations are positive but the magnitude of correlations are small (less than 0.35)). On the other hand, the correlation of Δ creativity and Δ word count may be due to two reasons. One possible explanation of this behavior explained in past literature [2] is that longer increase in slogan size may indicate participants may have devoted more time to the primary task, rather than to the secondary activity of chatting, which maybe reflected by higher creativity rating. Another possible explanation maybe that judges simply use length of slogan as a proxy of quality.

Analysis of chat logs

So far we investigated text similarity measures, final outcome scores, and correlations between measures to gain insights into different types of interactions. However, by studying only surface level measures, we largely considered chat interaction as a black box. As one would expect, some chats were more productive than others. In this section, we try to unravel those chat interactions by doing a grounded analysis of chat logs, subjectively categorizing them and investigating if there are characteristics common to each category of chat.

In conditions 'Chat' and 'Discuss' combined, 60 participants chatted with their partners to create a final slogan. The chat logs provided insights into different interaction types that occur between participants. To study the chat logs, we looked at surface-level text features (like number of words) and also manually tagged each chat with grounded labels to understand how the participants interacted, as described below.

Finding productive chat logs

To dive deeper into the chat logs, we asked two researchers with experience in rating slogans to group chat logs into categories. The two researchers derived these categories based on how effective the chats were in coming up with a new idea based on initial ideas. They first went through all the chat logs and then discussed categories into which each log can be bucketed into. In determining the categories, an important criteria was how the chat affected the generation of final ideas. They found that in a few interactions; both members relied heavily on each other's input; this led to the 'Productive Chat' category. In a few interactions, one member could have come up with the final idea but the partner completely relied on it to generate their idea. These interactions were tagged as 'Chat dominated by one person's idea'. Finally, chats where participants never talked or discussed unrelated topics were categorized as 'Non-productive chat'. Grounded in the logs, the research team decided on three categories which seemed to define most conversations: a) Non-productive chat, b) Chat dominated by one person's idea, and c) Productive chat—define in more detail below. The team defined a common rubric and independently categorized each chat log into one of the three categories. The Cohen's kappa between the raters was 0.69 after initial assessment. For the handful of chat logs where the raters disagreed, they discussed their rationale and mutually decided on a final category.

a) **Non-productive chat**: For this category, the participants had a one sided conversation (no response from other person) or only discussed aspects unrelated to refining the content of the slogan (*e.g.*, below box). We assigned 8 out of 30 chat logs to this category (4 in Chat, 4 in Discuss). Table 2 shows this condition's fixation, mimicry, convergence, and creativity scores. As expected, teams with near zero interaction converged less. Below we give an example chat and how the slogans changed:

> Participant A initial slogan: A top tier bicycle that fulfills the need of thousands of urbanites. Use this bike-sharing service daily with an affordable membership. This sleek bike will fit your modern lifestyle!
> Participant B initial slogan: Whether you are a casual or a serious rider, this new electric bike is sure to satisfy you throughout your riding endeavors. Take part in the new generation of bikes and have yourself feeling electric. Be the gold medal to your pedal.
>
> **Chat Log:**
> Participant A: *I think we should combine your electric idea with my bike sharing idea. I also like your slogan.*
> Participant B: *I dont know how we can combine them when they are two different products being sold. Ahh actually there it is. Our final doesn't need to be the same does it?*
> Participant A: *I don't think so.*
>
> Participant A final slogan: Whether you are a casual rider or a busy urbanite, this top tier electric bicycle will suit your needs. Use this bike-sharing service whenever you like, or pay for an affordable membership. Take part in the new generation of bikes and have yourself feeling electric!
> Participant B final slogan: Not only is this bike-sharing service affordable, but we have now included an electric bike package. This bike has a range of 35 miles on a single charge! Use this service to get to work, meet with friends, get groceries, etc...

b) **Chat dominated by one person's idea**: For this category, one participants may have proposed a combined slogan with little to no contribution from the other person. We assigned 8 out of 30 chat logs to this category, six of which were in the Discuss condition. This condition had high convergence because one participant adopts the other participant's idea rather than iterating on a joint solution. For example:

> Participant A initial slogan: Bike into the future
> Participant B initial slogan: A burst of energy for your new street smart life style.
>
> **Chat Log:**
> Participant A: *I like your idea and a burst of energy is a good term*
> Participant B: *I like your idea too*
> Participant A: *A burst of energy for you street smart lifestyle to help you bike into the future?*
> Participant B: *perfection*
> Participant A: *it has a call to action*
> Participant B: *exactly*
> Participant A: *and some visual aspects to it*
> Participant B: *A brillant merge of ideas. brilliant*
> Participant A: *anything else we could add?*
> Participant B: *I think we are good. unless you want to*
>
> Participant A and B final slogan: A burst of energy for your street smart lifestyle to help you bike into the future!

c) **Productive chat**: In contrast, if both people contributed and refined the slogan such that it incorporated information beyond what was contained in the Expose or Explain conditions, we categorized the chat as *productive*. We found 14 out of 30 chat logs were in this category, 9 of which were from the Chat condition. The larger number of productive chats in the Chat condition could explain the lesser perceived diversity and higher enjoyment scores compared to Discuss. As expected, most teams mutually agreed on the same final slogan leading to high convergence. Comparatively, chats in this category had lower fixation (as partners do not just stick to their own idea) and higher average mimicry (as both partners mimic parts from the other person's slogan). For example:

Participant A initial slogan: The bike of the future is here now.
Participant B initial slogan: Experience the Tour de France right in your backyard. Get your bike today.

Chat Log:
Participant A: *I like the TDF idea but think we should mention the futuristic style*
Participant B: *yeah its def futuristic. maybe be like "what the pros wish they had". make it feel like its so new even they dont have it yet*
Participant A: *i like it. what about - What the Tour De France riders will ride in the future, right now!*
Participant B: *that might be too obvious that there may be reason they dont have it yet. what im kinda thinking right now: "All the speed of a Tour De France racer while gliding in comfort never seen before. Get the bike the pros wish they had today."*
Participant A: *i dig it. lets go with that.*

Participant A and B final slogan: All the speed of a Tour De France racer while gliding in comfort never seen before. Get the bike the pros wish they had today.

Chats in Discuss were more one-sided

Among the surface level features, we measure the number of back and forth chats, the total number of words in the chat and the participation ratio—the ratio of the number of chat statements by one person to her partner. We found that the average number of statements in Chat and Discuss were 9.2 and 7.5 but the difference was not statistically significant. The average number of words in the chat logs for both conditions was also the same (80 words). However, the average participation ratio for Discuss was only 0.49 compared to 0.71 for Chat (a statistically significant difference). A perfect ratio of 1 implies that both partners contributed an equal number of chats, while a ratio of 0 would indicate that only one person sent all chat messages. This difference in scores provided evidence that team chats in Discuss were often one-sided compared to Chat.

We notice that the differences between Chat and Discuss were evident both from the participation ratio and from the manual annotation of chat logs. The low average participation score for Discuss indicated that it had many chat logs with one person doing most of the talking. The annotations found that Discuss indeed had more teams being dominated by one individual, while the Chat condition had a larger proportion of teams productively chatting. This may be responsible for the low average helpfulness and enjoyment score for teams compared to Chat (Table 1). As the total task time was same across both conditions and a participant cannot finish the task until the timer runs out, it is unlikely that participants in Discuss did not chat as they wanted to finish the task sooner. One possible explanation of the difference between Discuss and Chat maybe that participants in Discuss had enough information from their partners about the task, which led them to directly propose joint solutions rather than discussing possibilities. The final creativity scores for the two conditions were similar.

Category	$Creativity_f$	Δ creativity	Convergence	Mimicry	Fixation
Non-productive	1.70	0.12	0.57	0.53	0.77
One-sided	1.87	0.32	0.90	0.59	0.57
Productive	2.00	0.32	0.98	0.69	0.67

Table 2. Differences between chat categories show productive chats have high mimicry, high fixation and a higher degree of convergence.

DISCUSSION

Our study explores how idea generation in online dyads is affected by different modes of collaboration: we compare an interactive form of collaboration (chat) with a more reflective form (idea-explanation). Here we unpack the key findings.

Why did Explain have the highest increase in creativity?

Several factors could explain the increase in creativity for the Explain condition compared to Chat. One possible explanation for the consistent increase in creativity for the Explain condition is that allowing people to self-explain helps their partner to understand their slogan and come up with a better solution of their own. On the other hand, when team members chat online, they may waste time establishing rapport and figuring out how to work together. The Discuss condition may be implementing a form of production blocking [16], as users have to process multiple types of information (ideas, explanations, and chats) as well as generate their own chat responses and new ideas.

In the Explain condition, participants may have devoted more time to the primary task, rather than to the secondary activity of interaction, which is indicated in our data by larger increase in word counts. Similar to before, using ANOVA and Tukey's test, we found that the differences in increase in word counts (Δ wc) is significant (p value 0.02) with Explain condition (15.3 words) registering the largest increase in slogan size followed by Discuss (μ 7.3 words), Expose (4.3 words) and Chat (2.6 words). We observe that the mean number of words in the final slogans were also higher in the Explain condition (μ 32 words) compared to Chat (μ 24 words), Expose (μ 22 words) and Discuss (μ 25 words) conditions. However, the differences in slogan size across conditions was not significant.Longer slogans are an indication that participants spent more time writing.

Secondly, the increased creativity for the Explain condition might be understood by the fact that only 47% chats were found to be productive—that is, where both participants contributed to the team task. The teams which were not productive had basically the same information as the Expose or Explain condition. In contrast, explaining one's idea gave a person an opportunity to not only critically think about their own idea but also read the explanation of their partner.

Another explanation is that providing explanations may have allowed the task to remain loosely coupled. Olson *et al.* [43] reviewed research on collocated and non-collocated synchronous group collaborations for teams working remotely. They found four factors to be key for effective work in teams—common ground, coupling of work, collaboration readiness, and collaboration technology readiness. By enabling chat, the activity naturally gravitated towards a more tightly coupled activity (which implicitly meant coming to some agreed upon slogan). This more tightly coupled activity (even if only implicitly so) required dyads to reach common ground first. Future work could focus on explicitly measuring the coupling of interactions.

Why were Chat teams more likely to converge?

One of the proposed benefits of brainstorming is that it promotes mutual inspiration. Hearing others' ideas should allow group members to explore new categories that otherwise might have been not explored. Furthermore, 'piggybacking' might occur where one builds ideas off another group member's idea. By measuring fixation, mimicry and convergence scores in the Chat and Discuss conditions, we can offer insights in regard to piggybacking.

We found that both participants submitted the same idea for nine teams in Chat and eight teams in Discuss. We verified this observation by qualitatively analyzing the chat logs and noticed that in many cases, users first agreed upon a shared slogan during the chat and then submitted it. Comparatively, the limited interaction allowed by showing a selected slogan (in Expose and Explain) and explaining ones slogan (in Explain) makes achieving convergence more difficult. This shows that if the aim of the creative exercise is to gain convergence on ideas, allowing members to chat helps. This seemingly expected observation can be important in selecting the type of interaction suitable for a task. For instance, if tasks require a joint creative submission by the team, the Expose or Explain conditions only may not be sufficient.

We observed that 27% of teams (in the category "Chat dominated by one partner's idea') seemed to rely solely on their partner's effort for the final submission. We caution readers that such mimicry or fixation behavior may also be affected by factors like a partner's expertise level and how easily partners communicate. It is also possible that some social loafing occurred in Chat where individuals gave less effort to the group due to diffused responsibility. While we chose slogan writing to demonstrate our results, the benefits of the Explain condition may extend beyond that particular domain.

Limitations

We selected the slogan design task because it is representative of problems which are open-ended, can benefit from different viewpoints, and can be completed within a short time. However, it is unclear if our results will generalize to more complex design problems [29]. Secondly, the motivations of paired crowd workers and real-world teams may differ; providing different incentives may motivate teams differently, and thus alter an intervention's effects. Our experiments assumed that both participants complete the task synchronously and finish it within a limited time interval. Although synchronous interaction is not a necessity for 'Explain' and 'Expose' conditions, it was required for dyads who chat. Hence, future work is required to establish generalizability of our findings to other contexts, domains, and timeframes. Our work assumed that workers have no prior knowledge of each others background or existing social ties; future work can explore how our results change when partners know each other or have existing social ties. Lastly, we studied how participants interacted within a limited time frame; the interaction effects may differ when given longer interaction times.

Future work

Our results for the Chat and Discuss conditions imply that teams often seek consensus. While this may often be necessary for teams to carry out goals, an overemphasis on consensus-seeking behavior can also result in premature convergence. A well-known example is "groupthink" [27] which can arise when groups place too much importance on attaining consensus and fail to debate important alternatives for fear of damaging group cohesion. Possible ways to address this include encouraging team diversity, bringing in differing perspectives, and promoting healthy debates and dissents [54]. Our study paired individuals randomly, but it could be interesting to see how different team formation strategies [1] could help to avoid premature convergence during chat.

Another possible extension of our work can be integrating it with the 'team dating' approach explored by Lykourentzou *et al.*[37], where people interact on brief tasks before working with a dedicated partner for longer, more complex tasks. Such team formation exercises can allow people to choose partners according to their own subjective preference, which when combined with idea-explanation can lead to further improvement in creative outcomes.

In future work, we will develop a workflow for different creative tasks. The workflow may involve first generating ideas individually and explaining them, then reading ideas and explanations from a partner to inspire new variants, and finally chatting with the partner to converge. We believe our results provide implications for the potential benefits of introducing indirect interactions among multiple participants in complex and more general tasks, and we hope more experimental research will be conducted in the future to carefully understand the effects of idea-explanation in various crowdsourcing contexts.

CONCLUSION

In this paper, we study how dyads can generate better ideas by drawing inspiration from their partners by comparing interactive collaboration (chat) with a more reflective form (idea-explanation). Our findings showed that creative outcomes improve for users who draw inspiration from their partner's idea-explanation compared to either just seeing the partner's idea or chatting with that partner. We also show that participants who chatted were more likely to reach convergence on their final outcome. These observations indicate an alternative way to organize creative tasks in dyads: people solving tasks independently, practitioners may systematically organize people to work in pairs and enable indirect interactions that may enhance the creative output. For tasks requiring agreement on final outcome, dyads can be allowed to chat. In some sense, our results suggest the promise and potential benefits of working in pairs.

REFERENCES

1. Faez Ahmed, John P. Dickerson, and Mark Fuge. 2017. Diverse Weighted Bipartite b-Matching. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. IJCAI, Melbourne, Australia, 35–41. DOI: http://dx.doi.org/10.24963/ijcai.2017/6

2. Faez Ahmed and Mark Fuge. 2017. Capturing winning ideas in online design communities. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, ACM, Portland, USA, 1675–1687.

3. Salvatore Andolina, Hendrik Schneider, Joel Chan, Khalil Klouche, Giulio Jacucci, and Steven Dow. 2017. Crowdboard: augmenting in-person idea generation with real-time crowds. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. ACM, ACM, New York, USA, 106–118.

4. Paul André, Robert E Kraut, and Aniket Kittur. 2014. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, USA, 139–148.

5. Claire E Ashton-James and Tanya L Chartrand. 2009. Social cues for creativity: The impact of behavioral mimicry on convergent and divergent thinking. *Journal of Experimental Social Psychology* 45, 4 (2009), 1036–1040.

6. Jonali Baruah and Paul B Paulus. 2016. The role of time and category relatedness in electronic brainstorming. *Small Group Research* 47, 3 (2016), 333–342.

7. Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, ACM, New York, USA, 33–42.

8. Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, ACM, New York, NY, USA, 333–342.

9. Cheryl A Bolstad and Mica R Endsley. 2003. Tools for supporting team collaboration. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. SAGE Publications Sage CA: Los Angeles, CA, 374–378.

10. Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. DOI: http://dx.doi.org/10.18653/v1/S17-2001

11. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 169–174. https://www.aclweb.org/anthology/D18-2029

12. Joel Chan, Steven Dang, and Steven P Dow. 2016. IdeaGens: enabling expert facilitation of crowd brainstorming. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, ACM, New York, NY, USA, 13–16.

13. Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 2334–2346.

14. Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, ACM, New York, NY, USA, 1139–1152.

15. Alan R. Dennis and Mike L. Williams. 2003. *Electronic brainstorming*. Oxford University Press, New York, NY, US, Chapter Electronic brainstorming: Theory, research, and future directions., 160–178. DOI: http://dx.doi.org/10.1093/acprof:oso/9780195147308.003.0008

16. Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53, 3 (1987), 497.

17. Mahmoud Dinar, Jami J Shah, Jonathan Cagan, Larry Leifer, Julie Linsey, Steven M Smith, and Noe Vargas Hernandez. 2015. Empirical studies of designer thinking: past, present, and future. *Journal of Mechanical Design* 137, 2 (2015), 021101.

18. Steven P Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel L Schwartz, and Scott R Klemmer. 2012. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Design thinking research*. Springer, 47–70.

19. Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*. AAAI, Austin, Texas, USA, 32–41.

20. Michelle K Duffy, Jason D Shaw, and Eric M Stark. 2000. Performance and satisfaction in conflicted interdependent groups: When and how does self-esteem make a difference? *Academy of Management Journal* 43, 4 (2000), 772–782.

21. Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.

22. Nitesh Goyal, Gilly Leshed, Dan Cosley, and Susan R Fussell. 2014. Effects of implicit sharing in collaborative analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 129–138.

23. Terry Halfhill, Eric Sundstrom, Jessica Lahner, Wilma Calderone, and Tjai M Nielsen. 2005. Group personality composition and group effectiveness: An integrative review of empirical research. *Small group research* 36, 1 (2005), 83–105.

24. David A Harrison and Katherine J Klein. 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of management review* 32, 4 (2007), 1199–1228.

25. David A Harrison, Kenneth H Price, Joanne H Gavin, and Anna T Florey. 2002. Time, teams, and task performance: Changing effects of surface-and deep-level diversity on group functioning. *Academy of management journal* 45, 5 (2002), 1029–1045.

26. Sujin K Horwitz and Irwin B Horwitz. 2007. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management* 33, 6 (2007), 987–1015.

27. Irving Lester Janis. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes*. Vol. 349. Houghton Mifflin Boston.

28. David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies* 12, 1 (1991), 3–11.

29. James C Kaufman, John Baer, and others. 2005. *Creativity across domains: Faces of the muse*. Psychology Press.

30. Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55 (2004), 623–655.

31. Nicholas W Kohn, Paul B Paulus, and YunHee Choi. 2011. Building on the ideas of others: An examination of the idea combination process. *Journal of Experimental Social Psychology* 47, 3 (2011), 554–561.

32. Nicholas W Kohn and Steven M Smith. 2011. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology* 25, 3 (2011), 359–371.

33. Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham. 2013. Real-time Crowd Labeling for Deployable Activity Recognition. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1203–1212. DOI: **http://dx.doi.org/10.1145/2441776.2441912**

34. Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R Klemmer. 2010. Designing with interactive example galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2257–2266.

35. John M Levine and Richard L Moreland. 1990. Progress in small group research. *Annual review of psychology* 41, 1 (1990), 585–634.

36. Abraham S. Luchins and Edith Hirsch Luchins. 1959. *Rigidity of behavior: A variational approach to the effect of Einstellung.* Univer. Oregon Press, Oxford, England. xxv, 623–xxv, 623 pages.

37. Ioanna Lykourentzou, Shannon Wang, Robert E Kraut, and Steven P Dow. 2016. Team dating: A self-organized team formation strategy for collaborative crowdsourcing. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1243–1249.

38. Brian Mullen, Craig Johnson, and Eduardo Salas. 1991. Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and applied social psychology* 12, 1 (1991), 3–23.

39. Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL-HLT*. 568–578.

40. Bernard A Nijstad, Wolfgang Stroebe, and Hein FM Lodewijkx. 2002. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of experimental social psychology* 38, 6 (2002), 535–544.

41. Laura R Novick. 1988. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 3 (1988), 510.

42. Lynn R Offermann, James R Bailey, Nicholas L Vasilopoulos, Craig Seal, and Mary Sass. 2004. The relative contribution of emotional competence and cognitive ability to individual and team performance. *Human performance* 17, 2 (2004), 219–243.

43. Gary M Olson and Judith S Olson. 2000. Distance matters. *Human–computer interaction* 15, 2-3 (2000), 139–178.

44. Michelle A Pang and Carolyn C Seepersad. 2016. Crowdsourcing the Evaluation of Design Concepts With Empathic Priming. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, ASME, V007T06A004–V007T06A004.

45. Paul B Paulus, Nicholas W Kohn, Lauren E Arditti, and Runa M Korde. 2013. Understanding the group size effect in electronic brainstorming. *Small Group Research* 44, 3 (2013), 332–352.

46. Paul B Paulus and Bernard A Nijstad. 2003. *Group creativity: Innovation through collaboration*. Oxford University Press.

47. A Terry Purcell and John S Gero. 1996. Design and other types of fixation. *Design studies* 17, 4 (1996), 363–383.

48. Lauren E Scissors, Alastair J Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based CMC. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, ACM, New York, NY, USA, 277–280.

49. Garold Stasser, Susanne Abele, and Sandra Vaughan Parsons. 2012. Information flow and influence in collective choice. *Group Processes & Intergroup Relations* 15, 5 (2012), 619–635.

50. Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 459–468.

51. Eleni Triantafillou, Jamie Ryan Kiros, Raquel Urtasun, and Richard Zemel. 2016. Towards generalizable sentence embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. 239–248.

52. Hao-Chuan Wang, Chun-Yen Chang, and Tsai-Yen Li. 2008. Assessing creative problem-solving with automated text grading. *Computers & Education* 51, 4 (2008), 1450–1466.

53. Hao-Chuan Wang and Susan Fussell. 2010. Groups in groups: conversational similarity in online multicultural multiparty brainstorming. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, ACM, New York, NY, USA, 351–360.

54. Katherine Y Williams and Charles A O'Reilly III. 1998. DEMOGRAPHY AND. *Research in organizational behavior* 20 (1998), 77–140.

55. Teng Ye and Lionel P Robert Jr. 2017. Does collectivism inhibit individual creativity?: The effects of collectivism and perceived diversity on individual creativity and satisfaction in virtual ideation teams. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, ACM, New York, NY, USA, 2344–2358.

56. Doris Zahner, Jeffrey V Nickerson, Barbara Tversky, James E Corter, and Jing Ma. 2010. A fix for fixation? Rerepresenting and abstracting as creative processes in the design of information systems. *AI EDAM* 24, 2 (2010), 231–244.