Self-organization in online collaborative work settings

Ioanna Lykourentzou
Utrecht University, The Netherlands

Federica Lucia Vinella
Utrecht University, The Netherlands

Faez Ahmed
Massachusetts Institute of Technology, USA

Costas Papastathis
University of Peloponnese, Greece

Konstantinos Papangelis
Rochester Institute of Technology, USA

Vassilis-Javed Khan
Eindhoven University of Technology, The Netherlands

Judith Masthoff
Utrecht University, The Netherlands

Abstract
As the volume and complexity of distributed online work increases, collaboration among people who have never worked together in the past is becoming increasingly necessary. Recent research has proposed algorithms to maximize the performance of online collaborations by grouping workers in a top-down fashion and according to a set of predefined decision criteria. This approach often means that workers have little say in the collaboration formation process. Depriving users of control over whom they will work with can stifle creativity and initiative-taking, increase psychological discomfort, and, overall, result in less-than-optimal collaboration results—especially when the task concerned is open-ended, creative, and complex. In this work, we propose an alternative model, called Self-Organizing Pairs (SOPs), which relies on the crowd of online workers themselves to organize into effective work dyads. Supported but not guided by an algorithm, SOPs are a new human-centered computational structure, which enables participants to control, correct, and guide the output of their collaboration as a collective. Experimental results, comparing SOPs to two benchmarks that do not allow user agency, and on an iterative task of fictional story writing, reveal that participants in the SOPs condition produce creative outcomes of higher quality, and report higher satisfaction with their collaboration. Finally, we find that similarly to machine learning-based self-organization, human SOPs exhibit emergent collective properties, including the presence of an objective function and the tendency to form more distinct clusters of compatible collaborators.

Corresponding author:
Ioanna Lykourentzou, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands.
Email: i.lykourentzou@uu.nl

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
**Keywords**

Online collaborative work, distributed work, complex work, macrotask, self-organization

---

**Significance Statement**

Working in online platforms, where employers hire groups of workers on demand, is becoming increasingly common. Algorithms play a central role in coordinating this new labor model. Current algorithms organize people without giving them a say in who they will work with, when, or how. Can we give workers control while maintaining quality? In this paper, we propose a new approach, namely Self-Organizing Pairs (SOPs), which incorporates self-organization in algorithmic collaboration management, by facilitating workers to self-organize into effective pairs while being supported—but not guided—by an algorithm. We compared SOPs with two baselines that do not allow user agency, on a creative task of fictional story writing. Our findings indicate that the SOPs approach leads to creative outputs of higher quality and to higher collaboration satisfaction. Our results have scientific, applied, and societal implications. At a scientific level, we contribute to the development of collaborative work systems that are human-centered, non-intrusive and make workers an integral part of the algorithmic decision-making processes. At an applied level, SOPs can help existing online work platforms integrate the element of collaboration more systematically into their structure. Finally, this work contributes to the broader societal discussion around the future of work and ethical AI. Our approach, offering an alternative yet practical way to ensure a balanced coexistence between human workers and collaboration management algorithms, can contribute towards creating a future of online work that is more respectful and better uses both human and algorithmic capabilities.

---

**Introduction**

As online work increases in complexity, crowdsourcing research and practice looks into collaboration. Examples of problems where large-scale online collaborative work has proven valuable include scientific research and article authoring (Vaish et al., 2017), designing software prototypes (Retelny et al., 2014), writing stories (Kim et al., 2014, 2017), and collaborative idea generation (Siangliulue et al., 2016). The scale of crowdsourcing and its online nature, means that algorithms are often involved when it comes to deciding which crowd participants will work with whom. This approach is in contrast to face-to-face settings, where a manager or an expert user usually decides the members of a work group based on knowledge of their skills or of their past history working together (Gorla and Lam 2004).

Algorithm-mediated methods for group formation attempt to do the same at scale, either before the task starts, by assigning workers to groups based on a set of pre-calculated profile features (Rahman et al., 2019), or during the task, by changing the initial group syntheses to improve collaboration elements such as viewpoint diversity (Salehi and Bernstein, 2018) or interpersonal compatibility (Lykourentzou et al., 2016). Broadly speaking, these algorithms belong to the category of crowdsourcing management algorithms, the objective of which is to match persons to other persons or persons to tasks, and in this way to optimize the speed or efficiency of crowdsourced task production.

The problem with existing algorithms is that they tend to manage crowd collaborations in a top-down fashion, giving workers little or no say in who they will work with, for how long or how. For algorithms that assign individuals to groups before the task starts, the algorithm deduces the quality of the individuals’ interactions based on pre-calculated profiling features, such as expected performance or personality. Unfortunately, these features are often incomplete and subject to noise, given the sparsity of data in regards to, for example, complex skills. Elements such as interpersonal compatibility, which are inevitably only revealed after the group members have worked together, are also not taken into account. If the collaboration does not go as expected, workers cannot signal so, or change the course of the interaction. For algorithms that do change the group composition during the task, these changes happen without explicitly asking the workers or requesting their feedback. Workers receive a message that they have been placed with one or more collaborators and they are obliged to consent. User control is absent, with repercussions that include psychological discomfort (Rasmussen and Jeppesen, 2006) and less-than-optimal collaboration results (De Dreu and West, 2001). Recent studies (Lawler III and Worley, 2006; Retelny et al., 2017) suggest that monitoring the workflows too closely, for example, specifying precisely and in advance how the job will be performed or by whom, can stifle worker creativity and initiative-taking; two features that are necessary for the quality of creative and complex collaborative work. Not actively involving workers in the group
formation process but rather making decisions in a top-down manner, can also create power asymmetries and leave a disproportionate number of workers vulnerable to algorithmic biases and errors. These asymmetries are linked with the lack of transparency for which existing crowd work platforms are being criticized (Salehi et al., 2015; Rani and Furrer 2021) and to the broader discussion around algorithmic ethics (Mittelstadt et al., 2016) and accountability (Martin 2019; Wood et al., 2019a; Kellogg et al., 2020).

In parallel, a substantial body of literature in organizational psychology, management, and group research indicates the positive effects of agency and self-determination on group performance and worker well-being (Rasmussen and Jeppesen 2006; Van Mierlo et al. 2005, 2006). Performance-wise (Becker et al., 1998), allowing workers to co-design the task workflow—which for collaborative work means co-deciding the task roles and who works with whom—permits them to adapt to the task particularities, make better use of their expertise, creativity (Gilson et al., 2005; Gilson and Shalley 2004), and higher-level thinking (Amabile 1998). Collaborative work has also proven decisive for solving complex problems with no evident solution (Costa et al., 2018; Costa 2003). Agency increases a team’s ability to innovate even in the presence of disagreements; in fact, it is a prerequisite for the positive association between minority dissent (where a few group members disagree with the group decisions) and innovation (De Dreu and West 2001). In terms of worker well-being, agency leads to increased feelings of empowerment, intrinsic motivation, personal meaning, and reduces stress; these properties lead to increased teamwork satisfaction, cohesion (Carless and De Paola 2000; Gard et al., 2003; Man and Lam 2003), and commitment (Haas and Mortensen 2016; Kurtessis et al., 2017; Hackman and Oldham 1976). Further, when group members have a high degree of autonomy, they develop a sense of ownership and control over their own work and ideas, stimulating their creativity (Pratt and Jeffcutt 2009; Andriopoulos 2001). These positive effects of agency occur specifically for tasks with high interdependence among collaborators (Gilson and Shalley 2004; Langfred 2000b, 2005), such as complex, open-ended, non-easily-decomposable tasks that we investigate.

A problem therefore arises: (how) can we balance the necessity for algorithm-mediated collaboration formation, which is necessary given the scale of crowdsourcing, with the need to give each individual online worker a say in who they will work with to best accomplish the task? In this work we explore a new concept: **algorithm-assisted self-organization in collaborative crowd work settings**; an approach designed to provide online workers with the opportunity to choose their collaborators, and thus “guide” the algorithmic process of forming a collaboration. Self-organization is a management principle that has often been used in other settings, such as Open Source Software communities and agile corporate teams (Beck et al., 2001). To our knowledge, it has never been used in an online collaboration setting that requires the parallel involvement of an algorithm. **Self-Organizing Pairs (SOPs)**, which are the result of our approach, can be better understood through a metaphor from the machine learning field. In a similar way that Self-Organized Maps (SOMs) gradually re-arrange their input data to form highly coherent clusters based on Euclidean distances, people in the SOPs concept gradually discover the best collaborator to work with, based on reputation and personal experience of working together as the task progresses. An algorithm assists this process, by using people’s explicit collaborator preferences to form work pairs that are dictated not by an assumption of which collaborations could work, but based on which matches the workers indicate that will work. The fact that the algorithm does not rely on pre-established assumptions, for example, regarding which worker features to use to form the pair collaborations means that SOPs can be applied on a variety of complex online work tasks without the need to design a new sophisticated workflow every time the complex task changes, which is a problem often encountered by clients interested in crowdsourcing complex work (Sieg et al., 2010).

Self-organization in crowdsourcing is a new research line, and many alternatives can be explored regarding its implementation. In this paper we work with a collaborative-competitive setting, where the task is accomplished in discrete steps, called rounds. The specific task that we work with is collaborative short story writing, chosen as it is a complex task that does not require specialized expertise, does not have one evident solution, and can highly benefit from co-creation (Dobao and Blum, 2013). During each round, workers collaborate in pairs (herein after also dyads) to continue a main story. Dyads were chosen for this study as the essential instance of social group (Levine, 2003) exhibiting key collaborative online work processes, like coalition formation, power dynamics, reciprocity, and performance (Williams, 2010; Wang et al., 2013), linked to collaboration formation agency (Fauing et al., 2013; Corderoy et al., 2010; Hoegl and Parboteeah, 2006; Langfred, 2000a), which is the focus of this study’s framework. Dyads also exhibit a unique one-to-one capability to share and exchange creative ideas with fewer inhibitors of larger groups, such as social loafing, groupthink or production bottlenecks (Brajovich, 2003; Knights and Murray, 1994). At the end of each round workers decide as a collective, on a single “winning” story continuation using peer review. Before the next round starts, they can opt to stay with the same collaborator or to change, and an algorithm accommodates this process based on their choices. Gradually, round by round, workers get to explore the “space” of candidate collaborators to discover those with whom they might work best to win. We juxtapose this setting with one where the dyads remain fixed throughout the process,
without any choice for self-organization or agency. Results shed light into how people select their collaborators and how they gradually form desirable collaboration “clusters,” the effect that self-organization has on their collaboration, but also on their self-perceived effectiveness, sense of control and in general on the way that they decide.

The rest of this paper is organized as follows. In the next section we present related literature, including the limitations of current algorithm-driven group formation. We also summarize related work insights about self-organization from other domains, and present the hypotheses of this study. Next, we go through the study design, including the self-organizing pair framework and its supporting SOP algorithm. Then we describe the experimental process and the main experimental results. Finally, we discuss the study findings, with special emphasis on future directions given the novelty of this approach, and conclude the paper.

Related work

Collaboration formation algorithms: A tendency for top-down decision-making

One may distinguish two types of algorithms guiding the crowd collaboration formation process: (i) algorithms selecting which worker will work with whom before the task begins and (ii) algorithms that manage the collaboration formation after the task has begun and throughout its duration. Both algorithm types tend to make their decisions in a centralized, top-down fashion. The first type of algorithms views crowd collaboration as a mathematical optimization problem. Assuming a large pool of workers with known profiles (e.g., skill level) and a large pool of tasks, the objective of the algorithm is to match each task with a set of workers to accomplish the task optimally within given constraints, such as deadline, upper budget, or minimum quality. In this line, Rahman et al. (2019) propose an algorithm that utilizes affinity and upper critical mass to recommend tasks to sets of online workers, taking into account the aggregated worker skills and cost of effort. In addition, Liu et al. (2015) propose a task pricing algorithm that attempts to assemble a group of crowd workers to complete a given task with the lowest cost. Both these works rely on predictive learning algorithms, which make their collaboration formation choices based on a limited set of pre-calculated worker profiling features, without worker feedback. The risk of relying on such algorithms is to reduce workers to a set of dimensions that do not account for the collaborators’ evolution, intentionality, and needs (Faraj et al., 2018; Ananny, 2016), and therefore risk creating rigid, incomplete, and less-than-optimal group structures.

The second type of algorithms manage the crowd collaboration structures during the task. In this research direction, Salehi and Bernstein (2018) propose Hive, an algorithm that rotates workers across teams based on viewpoint diversity, to improve the quality of a creative design task. Workers are not asked whether they would like to switch collaborators. Although the use of the rotation algorithm did produce higher quality task results, the authors acknowledge that forcing participants to work with specific people led to discomfort, and that participants actually wished to prevent algorithm rotation decisions. Zhou et al. (2018) propose an algorithm based on multi-armed bandits with temporal constraints, which explores different team structures and timings to apply these structures. The algorithm explores various exploration-exploitation trade-offs and chooses, from a finite set of possible structural changes, which change it should make and when. In this case too, the algorithm is the driver and principal decision-maker behind the team composition. On the analogous subject of crowd-led authored content, Kim et al. (2017)’s Mechanical Novel provides Amazon Mechanical Turk crowd workers the opportunity to create short fiction stories in loops of reflection and revision, and in a manner that decentralizes the decision-making process much more than past systems. Further in this research line, Valentine et al. (2017) and Retelny et al. (2014) propose Foundry, a crowd management system that relies on a top-down algorithm to “assemble” workers into role-based teams. Workers can request a change in the initial team structures, but the final decision is taken in a hierarchical manner by a small number of expert workers and the task requester. Eventually, workers are notified as to which team they pertain. Although this system does incorporate worker feedback, it does so in the form of worker suggestions and not decisions. Finally, Kim et al. (2014) propose Ensemble, a system to create stories through the crowd. Ensemble is also coordinated in a top-down manner, with teams that feature story “leaders” directing a higher-level story vision, and workers materializing this vision into concrete story pieces. Workers do not get to decide on the final story, and their contribution is limited to proposing drafts, comments, and votes, that is, assisting the leader.

Although these past works, and especially Salehi and Bernstein (2018) and Kim et al. (2017) touch on the subject of user agency, the present study explores collaborative behavior and outcomes in a setting that incorporates user control systematically, as the utmost prominent aspect of the system design. Finally, Salehi et al. (2017) and Lykourentzou et al. (2017) both propose systems of automated team formation, which take worker feedback into account regarding the quality of past collaborations. In both these systems, the workers evaluate their peers after having worked with them, and the system uses these ratings to calculate an overall benefit function that drives team formation on a new, self-contained task. Our study shares similarities with these works in that it also actively requests worker feedback regarding past collaborators. It differs in
that the objective function is not decided a priori, but it is generated on-the-fly by the self-organization decisions of the user collective.

Latest research (Retelny et al., 2017) acknowledges that externally, for example, by an algorithm, predefined planning of the way a group will work or its structure, is not optimal, especially for open-ended, complex tasks. The reason is that such planning can inhibit workers from adapting to the needs of the complex problem they need to solve, in real-time. The authors suggest that complex problems require approaches that enable open-ended adaptation. This work is in line with prior research on accountable governance work models, which showcases that creativity is fostered when individuals and groups have relatively high autonomy in their everyday work processes, and a sense of ownership and control over their own work and ideas (Andriopoulos, 2001; Amabile, 2018). Allowing independence around work processes also enables workers to resolve and adapt to problems thus better utilizing their expertise and creative thinking skills (Amabile, 1998).

Self-Organizing Pairs (SOPs) is an approach in this direction. In contrast to the centralized manner of organizing crowd work, the SOPs method enables workers to collectively decide, rather than suggest, on the best course of action as the task progresses, flexibly adapting both the involved worker pair structures and the output solution as the task unfolds.

Towards giving workers collaboration formation agency

User agency in online collaboration settings has been studied under at least two major approaches in the literature: as a product of direct user negotiation (or reciprocal agreement), and as a product of algorithmic mediation. The first approach demands mutual agreement among the workers before forming the collaborative structures. The second approach uses an algorithm to mediate this process and determine the work groups based on preferences. The problem of agency as the product of user negotiation has been investigated through agent-based modeling by research like the one by Guimera et al. (2005). Their work simulates the emergence of collaboration networks in creative enterprises based on the users’ propensity to collaborate under multiple constraints (team size, the fraction of newcomers in new productions, and incumbents’ tendency to repeat previous collaborations). The explicit intent to remain in a collaboration—ergo, their direct negotiation—is part of a study on self-assembled teams by Zhu et al. (2013). The study enabled online gamers to join or to leave virtual teams across some period. The players could only join teams sequentially, and their decision to remain in those was determined by whether (a) they played together synchronously, (b) the team did not change in size during the cooperation, and (c) the team became inactive for longer than 30 min (after which it was dismantled).

The study of Tacadao and Toledo (2015) observed the emergence of group self-assembly, under the different approach of algorithmic mediation. The model in Tacadao and Toledo (2015) is designed for collaborative learning scenarios where the cohorts produced are evaluated on the parameter constraints and the number of collaborator preferences they satisfy. Finally, the work of Meulbroek et al. (2019) studies algorithm-supported matchmaking in student teams, where the authors developed a system based on the CATME algorithm determining students ranking preferences. In our study, the algorithm supports the users’ choice whilst easing the complexity of the negotiation (e.g., cognitive overload, group size), which could slow down the task’s execution and deplete the users’ working memory resources. However, in contrast to the studies mentioned above that focus on simulations and non-work-related tasks, we propose a solution for self-organization with real users in an online collaborative work scenario.

On the particular area of online system design, research has only recently started exploring the perceptions of users when it comes to choosing their collaborator, if they are given the choice, (Gómez-Zará et al. 2019a) examine how people search for and how they choose who they would like to work with in online platforms. Their research indicates that users search for collaborators based on competence, common values, similarity in social skills and creativity levels, and prior familiarity. They also find that users eventually choose people who are well-connected, with many prior collaborators. Their study concludes that future systems should be hybrid, augmenting user agency with algorithms. In an educational setting, Jahanbakhsh et al. (2017) examine users’ perceptions regarding automated team formation. Their findings reveal that although users valued the rational basis of using an algorithm to form teams, they did identify mismatches between their preferred criteria and those of the algorithm, and expressed the need for having a say in the process. This study too recommends giving users more agency in the selection of their collaborators, and advocates for constrained self-formation, in line with earlier works in the educational domain (Bacon et al., 1999).

The notion of self-organization

Self-organization is a notion antithetical to centralization. It has been studied in various settings and from different perspectives, from complex systems (Heylighen 2008), agent-based simulations (Serugendo et al., 2003; Gustafsson and Sternad 2010), and machine learning (such as Self-Organized Maps) to the organization of human social networks and communities. Concerning online work, self-organization is often used to describe the governance of software development groups, within a company (Beck et al., 2001).
or in Open Source Software Development (Karatzogianni and Michaelides 2009), where it helps teams cope with the increased dynamism, resource control autonomy, and decentralization, which are inherent in today’s globalized environments (Di Marzo Serugendo et al., 2004).

Our previous work (Lykourentzou et al., 2016) explores the dynamic formation of crowd collaborations without prior knowledge of worker profiles, in an approach called team dating. The idea behind team dating is that task authors delegate team building to the crowd workers themselves and ask them to try out different candidate co-workers, evaluate them, indicate those that they like working with, and then make crowd teams based on these indications. Rokicki et al. (2015) also discuss the crowd’s self-organization, by studying various collaborative structures including balanced teams, self-organizing teams (built upon one worker as the first administrator, accepting/denyng the contribution of other members), and a combination of team and individual strategies. They find that teams outperform individuals at task annotation without impacting the end product’s quality. In this paper we explore the effects that self-organized dyadic collaboration formation has on the individual’s sense of entitlement, reward, and creative outreach, without the presence of a single-handed administrator to moderate and steer the group towards effectiveness.

Exploration-exploitation is another notion closely linked with self-organization, albeit with a different meaning than in this paper. Kamble et al. (2018) consider exploration-exploitation trade-offs in labor platforms where flash teams and on-demand tasks can be improved by the assistance of a matching algorithm modeled on a binary classification of the agents in a pairwise fashion. The exploration herein is defined as the learning performance of untested teams against the exploitation of repeating previously tested teams. The model relies on known and unknown features for near-optimal matching in consideration to the population distribution (a priori knowledge) and the payoff structure under the aggregated performance objective with the lowest regret. Unlike the aforementioned study, we present a self-organized collaboration formation framework which does not reduce the workers to specific terms (analogous to binomial label classification problems) and does not preclude existing knowledge of the collaborators’ performance prior to starting the work. More so, the intensification and diversification of the strategy is not led centrally by a coordinating algorithm, but defined by the worker’s initiative only, as we allow the collaborators to guide the systemic changes occurring across the work phases. Feng et al. (2019) work with a mathematical model of social team-building optimization (STBO) based on swarm intelligence theory. Here, the team-building phases lead to a converging point dependent on the exploration of new solutions and the exploitation of already visited neighborhoods considering the group’s energy and entropy. This algorithm greatly depends on a defined social hierarchy which is not a prerequisite in the model we propose, enabling the process of collaboration formation to functionally generate a hierarchy as a result. Finally, in Zhou et al. (2018) exploration-exploitation means changing or keeping the team structure, as decided by an algorithm based on a finite set of (five) decision elements. In our case it means changing one’s collaborator or staying with the same, as decided by the users based on an unknown set of decision elements, constrained only by the number of dynamics cues the users can process through their text-based interactions. Humans in this case take up the role of exploring the decision space, instead of an algorithm.

Lessons learned from the literature and contributions of this work

As we saw in the previous sections, current approaches in algorithm-mediated large-scale collaboration formation are primarily top-down. They afford workers little freedom regarding who they will work with, or concerning the group structures and the collective task outcome. Related literature highlights that top-down decision-making works satisfactorily for well-defined tasks with few knowledge interdependencies, which can be straightforwardly decomposed to microtask level. However, it is less appropriate for ill-defined, complex, and creative tasks, such as innovation generation or creative writing. Bottom-up approaches have been found to improve group performance in such tasks because they motivate workers to own and take responsibility for the creative process and its outcome, adapt to the task, and make better use of their expertise, creativity, and higher-level thinking. Aiming to address the above, in this work we will explore for the first time a new human-centered computational structure, namely, Self-Organizing Pairs (SOPs). The SOPs structure does not just give users agency over who they will work with or help them form these collaborations at scale using an algorithm. It also enables users to control, correct, and guide the task output solution as a collective by competitively filtering out weak candidate solutions and selecting the most promising one as the task progresses. Through the self-organization, supported but not guided by an algorithm, users gradually build a consensus-based and community-approved global solution.

Methodology

Reward

To motivate users to switch collaborators if needed, the workflow includes an element of competition in the form of a bonus payment (for paid workers) or an increase of the obtained score (for non-paid users). Every time a worker pair wins, its members gain an extra amount equal to the base pay (for paid participants) or the base score (for non-paid participants). In line with the latest recommendations
for academic requesters regarding fair payment, the base pay for Amazon Mechanical Turk (AMT) workers was 5 Euro for a total task time of approximately 30 min. Given the three rounds (three chances to win), worker remuneration could reach a maximum of 20 Euro. For non-paid participants, the monetary payment was transformed to a base score; these participants would get a base score of 5 points for participating in the task and 5 more points every time their pair won. As a further element of driving competition among non-paid participants, a leaderboard was shown at the end of the task, illustrating their placement relative to the other participants. As we discuss later in the results and in the Discussion section, the reward level may affect the collaborator selection decisions, and consequently the objective function of SOPs as a collective.

**Group formation requirements**

Group size is critical to performance for creative tasks. Research shows that the number of creative ideas per person increases as group size decreases (Bouchard Jr and Hare, 1970; Renzulli et al., 1974). Thornburg (1991) further shows that a group’s Creative Production Percent (the percent performance of a group compared to the performance of an individual) improves as group size decreases until it reaches its peak at group size of two, that is, dyads. The reason is that dyads have a unique one-to-one capability to share and exchange ideas. At the same time, the inhibitors that typically occur in larger groups, like social loafing, group-think, and production bottlenecks, are less likely to occur in these groups (Brajkovich, 2003; Knights and Murray, 1994).

Finally, dyad interactions permit observing key group processes like coalition formation, inclusion/exclusion, power balances and imbalances, leadership and followership, cohesiveness, and performance (Williams, 2010), linked with expressions of collaboration agency and autonomy in various collaborative settings (Fausing et al., 2013; Corderly et al., 2010; Hoegl and Parboteeah, 2006; Langfred, 2000a). For the above reasons, dyads have been extensively used in crowd and social network research (Miller et al., 2014; Chikersal et al., 2017; Ahmed et al., 2019; Lykourentzou et al., 2016, 2017; Huang and Fu, 2013; Rivera et al., 2010). Taking the above into account, in this study we work with dyads. Nevertheless, this design decision also limits the scope of the study to pair work interactions. In the Discussion section, we elaborate on these limitations and how our model can be extended to accommodate larger groups.

Another parameter to decide is the batch size, that is, how many people will be recruited for a single SOPs lifecycle. This needs to be an even number so that all participants find a pair to work with. The larger the batch size, the more options a participant has in selecting collaborators. However, too large a batch also means that people cannot process and compare all candidate collaborators effectively, due to short-term memory limitations and time restrictions. With the above in mind, we opted for batches of 6–12 people. This allows for an adequate number of different dyadic collaboration formations while keeping the cognitive load of processing multiple user profiles manageable (Knijnenburg et al., 2012; Bollen et al., 2010).

**Task description**

**Task requirements.** In defining the appropriate task for the study, we took into account a number of requirements. First, we needed a task that involves complex, open-ended work for which no single solution is evident, cannot be easily decomposed to fixed workflow structures, and requires workers to maintain the global context and full semantic overview of the problem while iteratively refining it (Altshuller, 1999; Majchrzak and Malhotra, 2013). Recent crowdsourcing literature refers to these tasks as macrotasks (Schmitz and Lykourentzou, 2018; Lykourentzou et al., 2019), distinguishing them from microtask-based work. Examples of candidate macrotasks included brainstorming (Chan et al., 2016), writing (Kim et al., 2014, 2017), prototyping (Valentine et al., 2017), product development, innovation development (Kittur et al., 2019), open research (Vaish et al., 2017), formulating an R&D approach, and so on. These macrotasks require the combination of the diverse knowledge, skills and creativity of multiple individuals (Lykourentzou et al., 2019). As such, these tasks can benefit the most from the SOPs structure, the purpose of which is precisely to enable the continuous ad-hoc adaptation of the solution output and work processes to the task needs. On the other hand, tasks that are close-ended, those with known knowledge and skill interdependencies (Argote, 1982), or tasks for which a specific work process can be determined a priori (Okhuysen and Bechky, 2009) would not be appropriate candidates, as these can be optimally solved through workflow management and crowd coordination algorithms, like the ones described in the Related Work section. Furthermore, the task needed to adhere to three key criteria for an online creative work setting, involving people working online together for the first time: no requirement of prior expertise, short duration, and ability to express creativity (Dow et al., 2011).

The task that was selected to fulfill the above criteria is a creative writing challenge, inspired by the exquisite corpse method (Brotchie and Gooding, 1995), where participant pairs co-create a fictional story by gradually building on each others’ contributions, across multiple rounds. Creative writing tasks of the above type, can be used for applications such as rapid game scenario design (e.g., to provide more truthfulness and content to online gaming AI) or to generate content for the creative industries (film making, advertisement, etc.). In line with the SOPs framework, the task
allows for cycles of collaboration, where the worker dyads work internally to produce candidate story continuations, and competition, where the dyads compete for the single best story continuation through peer review. The pre-authored story used as input to the creative writing task of this study is the following:

At a restaurant, Mary receives an SMS and reads the following message: “Your life is in danger. Say nothing to anyone. You must leave the city immediately and never return. Repeat: say nothing.” Mary thinks for a second and then …

**Timing.** The proposed framework is designed to work with an ongoing flow of users joining the task at slightly different times, as it is typically the case when working with commercial crowdsourcing platforms. The system is programmed to account for a minimum threshold of registrations (between 8 and 12, depending on the flow of the workers) and a maximum waiting time, after which it redirects the workers to unique batches of experiments. By monitoring and assessing the registration flow of the workers across multiple trial runs, we were able to determine the average batch size for the experiments without encountering critical levels of delays that could overtake a large portion of the task. Even though the job fitted some of the characteristics of micro-tasks (real-time, short-termed, and unique), to be able to hire workers from the Amazon Mechanical Turk platform, its core is designed to be executed as a macrotask (complex, collaborative, and open-ended).

**Study setting**

Mainstream crowdsourcing platforms do not encourage collaboration and disallow the worker allocation into self-organized groups. For this reason, we designed a tailor-made framework and its supporting platform. The proposed work has been designed with the intention to address the individual’s ability to leverage collaboration with a certain degree of freedom and given creative agency as structural part of the collaboration process.

**Overview of the proposed self-organizing framework**

Figure 1 illustrates the functionality of the proposed SOPs framework. SOPs are designed to function on the basis of a collaborative-competitive setting, completed across various rounds. During each round (collaboration phase) workers form pairs and work with their collaborators to progress a creative task, which in our case is the continuation of a—pre-authored and same for all worker pairs—short fictional story. At the end of each round (competition phase), users employ peer review to vote for their favorite story continuation, without the possibility to vote for their own pair’s story. The winning story is appended to the main story, the winning pair is announced and its members receive an award, which for paid crowd workers is monetary and for non-paid volunteer participants is in the form of a score. Then, before a new round starts, users decide whether they wish to continue with their previous collaborator, or change. The SOPs algorithm forms the worker pairs based on these choices, facilitating self-organization. The process repeats for three rounds.
formation decisions, forming new or old pairs and continuing the main story as it was formed in the previous round. At the end of the final round, users are presented with the outcome of their collective work in the form of a finalized main story. Users are also presented with a ranked list illustrating the number of times each user won in descending order. Finally, users fill in a questionnaire on their experience.

Task initialization: recruiting participants. Users register to our experimental platform in two ways. In the case of paid crowd workers, they enter with the credentials of the crowdsourcing platform used to hire them, in order to facilitate their automatic payment once the task finishes. In the case of volunteer participants, they register with a unique identification number. For each experiment of our study, once the desired batch of people has arrived, the experimental platform stops hiring people, and those registered are moved to the next step. From that step onward, the system is synchronized, meaning that all workers are moved from one step to the other after a specific amount of time has elapsed. Users are always shown the remaining time, the round that they are currently in and the amount they have won so far on the top of their screen.

Setting up: Instructions, demographics, and Individual creative work sample. Workers are presented with the task instructions, which briefly present the creative task, its goal and their reward upon completion. This stage takes just over a minute, and users are given the following instructions:

1. The task: Users are instructed to work in pairs to continue a short story in English. They are informed that the task has three rounds and that in each round the story that gets the most votes wins and is appended to the main story. This extended main story will be then shown as a prompt to all participants and a new round will begin.

2. The goal: Users are instructed that their goal is to be in the winning pair. They are informed that in each round the system will automatically match them with the same or with another collaborator. They are prompted to do their best so that their dyad’s story gets the most votes.

3. The reward: Users are explained that they get a [Base reward2] for participating in the task and an extra [Base reward] every time they are in a winning dyad (3 chances for this). They are also told that their maximum gain is [4 × Base reward].

Next, users are asked to fill in a short questionnaire about their demographic information, namely: (i) gender, (ii) age, (iii) ethnicity, (iv) education level, (v) employment status, (vi) prior experience in a creative task like the one they are about to work on, and (vii) self-perceived creativity levels. To assess the creative self-efficacy, we used the eight-item scale from Carmeli and Schaubroeck (2007) and Chen et al. (2001). In our experiments, this stage takes less than a minute for completion.

Creative writing: Sample story writing. Then, each user is presented with the start of a pre-authored fictional story (same for all users), and is asked to write a brief continuation for it. We use this input as a sample of the quality of the individual’s work (“writing sample”) in two ways. First, we add it to their profile, visible to all users of the batch, so that they can themselves determine that individual’s writing skills. Second, we also evaluate it separately, using an external crowd, for comparison purposes during our results’ analysis. Users have three minutes to complete the individual writing sample stage.

Self-organization decision: Collaborator selection. Next, users are moved to the collaborator selection step, illustrated in Figure 2. Here, they will select their preferred collaborator(s) from the full list of user profiles of the batch. Users can see each others’ profiles, where each profile contains the following information about them: (i) username, (ii) demographic information, and (iii) writing sample. They can also see the (iv) average rating each candidate collaborator has received by the people he/she previously worked with (“others’ rating”) and (v) rating the person looking at the profile page may have given to that particular candidate collaborator if they have already worked together in the past (i.e., “own rating”). Note that items (iv) and (v) are only shown from the second round onward after users have already worked together at least once (Figure 3).

In the collaborator selection stage, users will be asked whether they want to work with the same collaborator or not. Users are also asked to indicate up to two other candidate collaborators to work with. These latter choices are useful to the system for two reasons: (i) in case the user indicated that they do want to work with their previous collaborator, but that person is unavailable, or (ii) in case the user indicated that they no longer want to work with their previous collaborator. The SOPs algorithm will use these choices to construct a “preferences matrix” and form the worker pairs of the next round.

The collaborator selection stage is a critical step in self-organization. It demands users to quickly assess multiple sources of information across multiple users, and balance potentially conflicting candidate decisions: for example, the psychological safety of working with a person similar to them (McPherson et al., 2001), versus the choice of choosing a highly rated person with whom they might not have a lot in common. In the next rounds, when the information available to the users for making a choice increases, users will also need to individually assess their relative gain from continuing with the same collaborator.
lower communication overhead and potential presence of transactive memory (Hollingshead and Brandon, 2003), since the group has learned to work together) versus the risk of losing the chance to work with a new collaborator (for example, a previous round winner) who could potentially increase their chances of writing that round’s winning story. Users are given 2 minutes to choose their preferred collaborators.

Figure 2. Self-organization—Round 1. Self-organization takes place during the collaborator selection phase. Participants vote for their preferred collaborators based on the profiles of the latter (consisting of demographics, writing experience, creativity level and sample story), and the SOPs algorithm uses these votes to form the worker pairs. In the next rounds, participants also see the average rating of each person, as well as their own rating for that person (if that exists).

Figure 3. Participant profiles as seen by other users in our experimental platform interface—Collaborator selection phase, Round 1. In the next rounds, participants also have the choice to indicate if they wish to stay with their previous collaborator or not.
**Collaboration and internal evaluation phase.** As soon as the algorithm has placed users in work pairs based on their indicated preferences, each pair is moved to an online synchronous collaboration space, with a text writing area and capability to chat. The software Etherpad (Erdal and Seferoglu, 2017) was used to facilitate this kind of synchronous collaboration. The software automatically highlights each user’s input in a different color, so that the pairs can track who wrote what. The back-end of the system saves the chat and story progression. Figure 4 shows an example collaboration between two people, Kristy and Peter.

In the collaboration phase, each pair is instructed to continue the story so far (“main story”). In the first round, the main story is simply the initial pre-authored story presented to the users at the individual writing sample stage. In the next round, the main story will gradually increase, since after every round the winning pair’s story will be appended to it. The work pairs are free to discuss and collaborate to continue the main story, in any way that they like. This allows us to observe different group dynamics and interaction patterns, work and collaboration strategies, creativity patterns, etc. Each collaboration round lasts for four minutes. Thirty seconds before time is up, the workers can see which worker pair is leading the group. The story will gradually increase, since after every round the winning pair’s story will be appended to it. The work pairs are free to discuss and collaborate to continue the main story, in any way that they like. This allows us to observe different group dynamics and interaction patterns, work and collaboration strategies, creativity patterns, etc. Each collaboration round lasts for four minutes. Thirty seconds before time is up, the users also see a reminder to wrap up their story.

Once time is up, the worker pairs are asked to evaluate one another on three axes using a Likert scale of 1–5. (i) Skillfulness (“How skillful was [collaborator’s username] in continuing the story?”), (ii) Collaboration ability (“How good is [collaborator’s username] as a collaborator?”), and (iii) Helpfulness (“[collaborator’s username] comments were helpful”). Each group member is also asked to assess their own helpfulness level (“My ideas and comments were helpful”) on a Likert scale of 1–5. Finally, users are asked to assess the number of core competencies they noticed having in common with their collaborator (“[collaborator’s username] and I were similar in:"), with four possible options (multiple or none can apply): (i) task commitment (“Commitment to working hard on this task”), (ii) work strategy (“How we think the work should be done”), (iii) Skill similarity on task (“General abilities to do a task like this”), and (iv) Personal values (“Personal values”). These ratings are used to enrich the profile of each user, both in terms of the “other’s ratings” (average rating by previous collaborators) and in terms of the “own ratings” (of the person looking at that user’s profile), as explained earlier. The members of the worker pair have half a minute to complete their evaluation of one another.

We determined the timeline of the experiments after several experimental trials with multiple combinations of time slots. When adjusting these time slots we also took into consideration the AMT outsourcing model, which favors micro-tasks. Although extending the time for each phase of the task could have been beneficial to some workers, we noted that most were able to produce their judgment within the given time. Batch sizes did not differ greatly between experiments and the stories that needed to be voted on by each worker were no more than three at a time and considerably short in length. With batches significantly larger than what we used in this study, lengthier time slots would have been even more so applicable. We discuss further the scalability of the system in the Discussion section.

**Competition: Voting for best story and presenting the winning pair.** After evaluating their collaborator for that round, each individual user votes for their preferred story continuation, among the $S−1$ candidate continuations, $S$ being the total number of worker pairs from the previous round (users cannot see or vote for their own team’s continuation). In voting for the best story, users can see which worker pair (i.e., which two usernames) produced which story continuation. Users have one and a half minute to read and decide on their preferred continuation. Once the time is up, the story with the most votes (“winning story”) is presented to them, along with the usernames of the two members of the winning pair. The profile of winning pair members is updated so that the bonus amount for winning is added to their individual total earned reward. Presenting the winning story before users are asked to make a decision on their collaborator of the next round is important to give users an overview of their results so far, and reinforces the
competitive element of this phase of the framework. From a task point-of-view, the story peer assessment at this stage allows for a collective decision to emerge regarding the outcome of the task, that is, users collectively have full control over the task result. Peer review is also a proven way of incorporating quality assurance during the task (Whiting et al., 2017). Alternative ways of evaluating the team result after each collaboration round can be envisioned and they are relatively straightforward to incorporate, without affecting the core of the proposed system. These ways include assessment by an external crowd or by one person, such as the client who commissioned the task.

Next, and assuming the predetermined total number of rounds is not over, users return to the collaborator selection stage. As explained in the “Collaborator selection” section, here they must decide whether they want to continue with the same collaborator as in the previous round, or whether they want to change. In both cases, they are also asked to indicate up to two additional candidate collaborators, from the full candidate collaborator profile list, for the algorithm to use either in case it cannot accommodate their first choice (if they wanted to stay with their previous collaborator), or for the algorithm to use to match them with a suitable alternative collaborator (in case they wanted to change).

The cycle of self-organization-collaboration-competition continues, with the main story gradually increasing in length as more and more continuations are appended to it. After a number of rounds, which for the purposes of this study is set to three, users see the final story, the final user ranking (in a descending order based on the number of times a user has been a member of a winning pair), and a final questionnaire about their overall experience. Once they fill in this questionnaire, users are redirected to the crowdsourcing platform and receive their payment.

**Self-Organization algorithm**

The SOPs algorithm is one of the first to maximize user agency in algorithm-based crowd pair formation. Its aim is to assist but not dictate the self-coordination process, matching users with those collaborators that they mostly prefer working with. The algorithm receives as input the individual user profile ratings, encoded in a tabular form as follows. Assume user $A$ has worked with user $C$ in the previous round. If a user $A$ indicates that he/she wants to continue working with $C$ in the next round, then $A \rightarrow C = 3$. If user $A$ indicates that he/she does not want to continue working with $C$, then $A \rightarrow C = 0$. User $A$ can also select from the list of other candidate collaborators, with whom he/she has not worked in the previous round. Assume this list contains users $B$ and $D$, and assume the user indicates $B$ as a preferred collaborator for the next round, and does not indicate $D$. Then $A \rightarrow B = 2$ and $A \rightarrow D = 1$. In brief, a user’s preference is given the highest weight to those collaborators that the user has worked with and wants to continue working with, and the lowest weight is given to those that the user has worked with but does not want to continue working with. The aforementioned possible user-to-user values are summarized in Table 1.

Using a user rating vector per user (i.e., two per team) as shown in Figure 5(a), the algorithm then constructs a complete graph (“affinity graph”) with candidate pair members as nodes, and the average pairwise ratings between individual users as the edges (Figure 5(b)). Next, the algorithm identifies all possible candidate worker pairs, that is, all possible graph cuts of size two. Next, itanks the candidate worker pairs on a list based on their average pairwise rating (edge value) from the highest to the lowest (Figure 5(c)). In other words, the algorithm ranks the candidate worker pairs starting from those that want to work together again, continuing with those that have not worked together before but would like to, and ending with those that do not want to work together again. From this ranked list, the algorithm selects the first worker pair, and removes all other candidate pairs that contain the selected pair’s members (as one person can only be in one work pair at a time). Figure 5(d) shows the selected pair of $B$ and $C$ in green and removes gray nodes containing users $B$ or $D$ as options. The algorithm continues in this manner, until the list of candidate pairs is empty, and all users have been placed in a pair. In case of ties, the algorithm chooses randomly. The pseudo-code for this process is shown in Algorithm 1.

**Algorithm 1.** Self-organizing team formation algorithm. The algorithm creates the teams based on the ratings and preferences of the users regarding their candidate collaborators.

**Data:** Individual profile ratings  
**Result:** Final list $F$ of teams for next round

1. Create complete graph $G = (V, E)$: $V$: candidate team members, $E$: average pairwise ratings;
2. Find all possible graph cuts of size 2 (candidate teams), $→ C$;
3. Sort $C$ in descending order;
4. $F \leftarrow \emptyset$;
5. **While** $C \neq \emptyset$ **do**;
6. Pick first element $C_i = \langle x, y \rangle$ in $C$: $x, y \in V$;
7. $F = F \cup \{C_i\}$;
8. $C = C - C_i - \{C_i\}$: $x \in C_j$ | $y \in C_j$;
9. **return** $F$;

It must be noted at this point that the self-organizing algorithm described above is greedy, and although it is expected to perform sufficiently well for the scale of the experimental setting of the paper, it may not perform optimally in settings where the size of the candidate collaborator space is much larger and may include hundreds or even thousands...
of profiles. For these cases, one could examine the abstraction of the dyad matching problem to the maximum weight matching, stable matching or stable-roommate problems, and the subsequent adaptation of algorithms such as Edmonds (1965), Gale and Shapley (1962), or Irving (1985) to guide the pairings between the workers.

Next, we will describe the experimental conditions we designed to study our methods, which shed light on how dyadic self-organization affects quality of work and collaboration satisfaction.

**Experimental conditions**

For this study we work with three experimental conditions, examining the proposed approach and two benchmark conditions.

- **SOPs:** This condition stands for “Self-Organizing Pairs,” and studies the proposed approach of self-organization in online work pairs. People are given the choice to indicate their preferred collaborators, including the option to stay with the same person that they worked with in the previous round. The algorithm respects these choices and aims to place each person with the collaborator of their choice.

- **Placebo:** This is the first benchmark condition. This condition creates the illusion of agency, where participants believe that they can self-organize, but eventually they cannot do so. The workflow and experimental interface of this condition is identical to that of SOPs, that is, users do have access to the collaborator selection stage. The difference between the two conditions happens in the background, where participant choices at the collaborator selection phase are not taken into account by the algorithm. Instead, each participant is paired with a randomly allocated collaborator, and stay with them for the entire task.

- **No-Agency:** This is the second benchmark condition. It is made to resemble existing methods of placing people in ad-hoc online teams, where users lack agency over collaborator selection. It is also similar to the benchmarks used in the literature for testing online team formation from a worker crowd (see, for example, (Salehi et al., 2017)). In this condition, participants are paired randomly with one collaborator at the beginning of the task, and stay with them throughout the process, without the option to choose their collaborators (neither functional nor placebo). The difference with the Placebo condition is that the task workflow of the No-Agency condition skips the collaborator selection stage entirely. Since the teams do not break in either the Placebo or No-Agency conditions, these two benchmark conditions are also designed to capture collaboration continuity and transactive memory building benefits (Hollingshead and Brandon 2003) from staying with the same collaborator for longer periods of time.

**Participants.** A total of 140 people took part in a total of 18 experiments for this study. The study participants were recruited either as university students (68 participants) or as Amazon Mechanical Turk workers (72 participants), in batches of 4–12 people depending on availability. Eight of these dropped out due to Internet connection issues, resulting in a final total of 132 people who finished the experiment. The batches of participants belonging to these two different user groups (paid and volunteered) were managed in separate sessions and they were equally distributed between the conditions. The allocation of people to condition was made in a round robin manner to avoid biases due to participant type or batch size, resulting in six batches per condition. The total number of people who participated in

![Figure 5](image-url)

**Figure 5.** Steps of the algorithm’s operation. Nodes represent the participants, edges represent participant collaborator preferences (values 0–3, higher values mean higher collaborator preference). The algorithm (a) first constructs a complete bidirected graph comprising all user preferences, and then (b) it constructs the affinity graph comprising the average pairwise ratings. Next (c) it ranks all possible worker pairs (i.e., graph cuts) in descending order of collaboration preference and (d) respectively forms the pairs.
the Placebo condition was 52 and those participating in the SOPs condition was 48, and those participating in the No-Agency condition was 32.

To further exclude the possibility of confounding factors, we conducted a series of post-hoc checks. First, using the demographics information filled in by the participants in the beginning of the experiment (the Methodology section), the sample was controlled for statistically significant differences across the conditions in terms of demographics, namely, gender, age, ethnicity, education, employment status, prior experience, and self-perceived creativity. An analysis of variance (ANOVA) showed no significant differences across any of the aforementioned axes (all at \( p > .1 \)). A similar analysis also excluded any statistically significant differences in terms of individual writing skills between the three conditions, as evaluated by the external crowd evaluators who rated each participant’s individual writing sample, again in the beginning of the experiment, on the axes of grammar and syntax, interest, originality, plot structure, and overall impression of the story sample (\( p > .4 \) across all evaluation axes). Further, a regression analysis controlling for participant type showed no significant effect of this variable on the study results. Finally, we controlled on whether there was any difference in user perceptions of the benefit of agency across the three conditions. An analysis of variance comparing perceived collaborator selection usefulness, from the final questionnaire answers, showed no statistically significant difference across the conditions (\( p > .6 \)). These are the same criteria that, as we will see later, were used to evaluate the stories produced by the worker pairs during the collaboration. Having a sufficiently balanced sample across the three conditions, we proceed with the analysis of our results.

**Results**

We organize our results as follows. First, we look into the quality of the produced work by different worker pairs in the three conditions to investigate the question, “Did the pairs formed under the SOPs condition produce stories of higher quality than those of the Placebo and No-Agency benchmark conditions?” Second, we look into the quality of the collaboration to investigate the question, “Did participating in the SOPs condition enable participants to collaborate better and be more satisfied with the process of collaboration, compared to participants in the Placebo and No-Agency benchmark conditions?” After answering these key questions we look deeper into the mechanics of self-organization, examining two emergent patterns of self-organization, namely the presence of an objective function driving the collective, and network clustering phenomena.

**Work outcome quality: SOPs write stories of higher quality**

A total of 196 unique story continuations were produced by the teams. The final winning stories were 18. To evaluate the quality of these stories, we employed a crowd of external judges, hired through AMT. Each story continuation was evaluated by 10 AMT workers, on a ten-point Likert scale.
(1-10), and on five quality criteria: grammar and syntax (“How grammatically and syntactically correct is the story?” ranging from “Not correct” to “Very correct”), interest (“How interesting is the story?” ranging from “Not interesting” to “Very interesting”), originality (“How original is the story?” ranging from “Not original” to “Very original”), plot structure (“How good is the story plot?” ranging from “It doesn’t make sense” to “It flows nicely”), and overall impression (“Overall how much did you like the story?” ranging from “Not at all” to “Very much”). These criteria were selected as they are among the most frequently used by professional short story evaluators (Boden, 2004), and because they represent a balanced mix of both objective (grammar, plot structure) and subjective (interest, originality, overall impression) axes (Díaz Suarez, 2015).

An analysis of variance indicated that SOPs participants create stories of significantly higher quality than the benchmark condition pairs, across all five quality criteria, albeit with slightly different absolute value differences between the conditions, with $F_{\text{grammar}}(2, 1957) = 41.835, p < .001, \eta^2 = 0.041$, $F_{\text{interest}}(2, 1957) = 78.742, p < .001, \eta^2 = 0.074$, $F_{\text{originality}}(2, 1957) = 84.198, p < .001, \eta^2 = 0.079$, $F_{\text{plot}}(2, 1957) = 41.835, p < .001, \eta^2 = 0.041$, and $F_{\text{overall}}(2, 1957) = 99.847, p < .001, \eta^2 = 0.093$. Figure 6 illustrates these results.

A Tukey post-hoc test per quality axis also revealed that SOPs groups differed significantly from the other two benchmark condition pairs, across all five quality criteria (at $p < .001$), while the Placebo and No-Agency conditions differed significantly in terms of interest ($p < .05$), originality ($p < .001$) and plot ($p < .05$). A regression analysis analyzing the story continuation data, with round as a random effect, showed that the round does not account for the relationship between the higher performance of the SOPs condition compared to the others.

Finally, a Pearson correlation analysis between ratings for attribute versus the other (like grammar and overall) showed that the rater’s evaluations were significantly correlated ($p < 0.01$) with each other, indicating the presence of the “halo effect,” which is well-documented in many social judgment settings. The Halo effect implies that a rater’s judgments of one quality dimension tends to influence others, even in the presence of sufficient information to allow for an independent assessment of them (Nisbett and Wilson 1977; Saal et al., 1980; Woehr and Huffcutt 1994).

See Table 2 in the Appendix for detailed values.

Overall, the external ratings on multiple factors show that stories produced in the SOPs condition were better quality compared to the benchmark conditions. Next, we will discuss the perceived quality of their collaboration by participants in different conditions. Further visualizations and analyses of the method of estimation graphics (Ho et al., 2019) can be found in subsection “Estimation graphics” and Figure 11 of the Annex.

**Collaboration quality**

One of the major goals of team formation is to form effective collaborations. While effectiveness can be studied using many methods, we first focus on the perceived collaboration according to the individuals.

**People in the SOPs condition evaluate one another higher in collaboration, helpfulness and skillfulness.** On average, the work pair members in the SOPs condition rated each other significantly higher as collaborators (on a scale of 1–5) and in terms of how helpful they were, compared to the pairs formed under the Placebo and No-Agency benchmark conditions, with $F(2,211)_{\text{collab}} = 21.364, p < .001, \eta^2 = .168$ and $F(2,211)_{\text{help}} = 20.239, p < .001, \eta^2 = .161$, for the metrics of collaboration ability and helpfulness levels respectively. For each of these results, a Tukey post-hoc test was also run, revealing the presence of statistically significant differences between the SOPs condition and each of the benchmark conditions, as well as between the two benchmark conditions. Specifically, for the metric of helpfulness, the groups formed under the SOPs condition evaluated one another significantly higher compared to the groups formed under the other two benchmark conditions ($p < .001$ for each comparison), while there was also a statistically significant difference between the Placebo and No-Agency benchmark conditions (with $p < .05$). For the metric of collaboration ability, the groups formed under the SOPs condition also evaluated one another significantly higher compared to each of the groups formed under the other two conditions ($p < .001$ for each comparison). Here too, the post-hoc test revealed that the Placebo groups perceived their co-workers as more collaborative than the groups formed under the No-Agency condition ($p < .05$).

We find that the perception of helpfulness within the collaboration went both ways. Not only SOP members perceived their collaborator’s contribution as more helpful, but through the collaboration they also perceived their own contributions to the dyad as significantly more helpful. In contrast, participants in both the two benchmark conditions found that their ideas and contributions were not as helpful for their dyad, with $F(2,211)_{\text{lowHelp}} = 8.089, p < 0.001, \eta^2 = .071$. A Tukey post-hoc test showed that the aforementioned results were only significant as to the difference of the SOPs groups with the benchmark groups (with $p < .05$ between the SOPs and Placebo condition and $p < .001$ between the SOPs and No-Agency condition). Interestingly however, there was no statistically significant difference between the Placebo and No-Agency groups ($p = 0.262$) in terms of how they perceived their own helpfulness, although as we saw before, the two conditions did differ in how their groups perceived the helpfulness of their collaborator (No-Agency was lower).

We also observe that SOP members perceived their collaborators as significantly more skillful compared to
the perception that the members of the two benchmark conditions have of their collaborators’ skills, with $F(2,211)_{\text{skill}} = 15.366, p < 0.001, \eta^2 = .127$. A Tukey post-hoc test revealed that these differences are significant among all three groups, with $p < 0.05$ between the SOPs and Placebo conditions, $p < 0.001$ between the SOPs and No-Agency condition and $p < 0.05$ between the Placebo and No-Agency condition. Interestingly the aforementioned higher perception of skillfulness is not because SOPs members are indeed more skillful; in fact, as also mentioned above, participants in the three conditions do not differ statistically in terms of skillfulness as evaluated by external evaluators on their individual writing samples. Previous research (Hansen et al., 2002) demonstrates that when people are more satisfied by their collaboration, then they tend to think more highly of their peer, thus being more prone to associate affective trust to positive expectation about belonging to that team. For similar reasons, group cohesiveness can positively affect the perception of satisfaction and group performance. We note that from the three conditions, the No-Agency benchmark condition, that is, the one where people were not given any (not even a placebo) option to choose their collaborator was the one with the lowest intra-team evaluations in terms of all four axes of collaboration, helpfulness and skillfulness. These results, summarized in Figure 7, indicate that the teams formed under the SOPs condition are more satisfied during their collaboration, and able to collaborate and help each other more, despite not being objectively more skillful than the individuals of the benchmark condition teams. Further visualizations and analyses of the collaboration quality data using the method of estimation graphics (Ho et al., 2019) can be found in subsection “Estimation graphics” and Figure 12 of the Annex.

**SOPs members are more aligned in personal values, skill similarity, and, how the work should be done.** During their peer evaluations, the participants of each work pair also indicated which, if any, competencies they had in common with their collaborator. As explained in detail in the “Collaboration and internal evaluation” section, they could report a common sense of commitment to the task, work strategy similarity, skill similarity, and/or common personal values. Figure 8 summarizes the percentages of common items reported by the groups across the three conditions. Further visualizations and analyses of the common competencies data using the method of estimation graphics (Ho et al., 2019) can be found in subsection “Estimation graphics” and Figure 13 of the Annex.

A chi-square test of independence was performed to examine the relation between condition and number of work style items that the teams reported having in common. The relation between these variables was significant, $\chi^2(8, 215) = 32.29, p < .001$, with SOPs members reporting to have more items in common with their collaborator compared to both benchmark condition pairs. Post-hoc tests also showed that the pairs differed significantly in how similar they reported to be in terms of their personal values ($\chi^2(2, 215) = 21.22, p < .001$), skill similarity ($\chi^2(2, 215) = 10.48, p < .05$), and work strategy ($\chi^2(2, 215) = 28.72, p < .001$), while they did not differ significantly in their commitment to the task.

This result indicates that when people’s choice of a collaborator (SOPs condition) is honored, then they tend to pair with collaborators with whom they share values, work practices and skills, confirming prior literature (Tenney et al., 2009).

**SOPs groups write shorter stories and make more equal contributions.** The collaborations assembled under the SOPs condition produced shorter story texts ($m_{\text{expad}} = 234$ characters, SE = 13), compared to both the Placebo ($m_{\text{expad}} = 320$ characters, SE = 16), and the No-Agency condition teams ($m_{\text{expad}} = 556$ characters, SE = 88), with $F(2,212) = 17.395, p < 0.001, \eta^2 = .141$.

A Tukey post-hoc test revealed that these differences were significant among all three conditions, with $p < 0.05$. One explanation for this could be that the benchmark condition collaborators know each other better, since they have worked together in the past rounds, whereas the same is not always true for SOPs collaborators, who may or may not have worked together in the past. If that is the case, then the chats between the three conditions could be expected to differ statistically in length, with the SOPs groups chatting more in an effort to establish a common ground in each round; therefore having less time to work on the actual task. However, the analysis of variance comparing the total chat length between the three conditions showed that there is no statistically significant difference.

Looking deeper into the process of story writing, we examine level of collaboration equality in the way that the participants of the three conditions produce their common story text. We start by measuring the metric of “turn-taking.” Turn-taking is a property of collaboration (Sacks et al., 1978) based on construction contributed which allows two or more entities to build a discourse from separate units. The metric was chosen for the evaluation of the given experiments as it considers the amount and the timing of the individual contribution towards the group work.

We measure turn-taking as follows. First we identify every text piece (every segment of text entered by the users) each pair member entered in the common text area, and the order in which this member entered this piece. This gives a sequenced order of contributions. For example, assume a pair consisting of person $A$ and person $B$ and a writing sequence of their collaboration to be $\langle ABABAAA \rangle$. We encode this sequence as $\langle -1,1,-2,2,-3,-4,-5 \rangle$, sum and
normalize it by the sequence length. The turn-taking was tracked by the back-end part of the system, which saved the final version of each story continuation when the writing phase ran out of time. The software used for hosting this kind of synchronous collaboration, called Etherpad (Erdal and Seferoglu, 2017), helped to automatically highlight the text in a different color for each user, meaning that user A could see his or her text in a different color than that of user B.

Every time the contribution of one pair member is followed ("matched") by the contribution of the other member, the value of this metric is equal to zero. The more person A dominates the writing process, the more negative values the metric receives. The more person B dominates the writing process, the more positive the metric becomes. Hence, values around zero indicate a balanced writing process in terms of turn-taking.

The analysis of the story writing processes for the two conditions, using a random allocation of pair members to position A and B of the metric, shows that the SOP groups have significantly more balanced text logs, in terms of turn-taking style ($m_{\text{turn}} = 0.01, \ SE = 0.051$) compared to the pairs of the Placebo ($m_{\text{turn}} = -0.22, \ SE = .054$) and No-Agency conditions ($m_{\text{turn}} = -0.50, \ SE = .091$), with $F(2, 212) = 14.515, p < 0.001, \eta^2 = 0.120$. A Tukey post-hoc test revealed that these mean differences were significant among all three conditions, at $p < .05$.

Another way to look into the degree of equality in the collaboration of the worker pairs is by using the Gini coefficient $G$. This metric, typically used to compare income (in)equality among countries, can also be used to examine the degree to which the collaborations are balanced or they tend to be dominated by a few individuals. The metric receives values between 0 and 1, with 0 corresponding to a fully balanced collaboration and 1 corresponding to a fully imbalanced one. As an example,
consider three users contributing 5, 5, and 10 text messages, respectively ($u_1 = 5$, $u_2 = 5$ and $u_3 = 10$). 50% of the messages are of size 5 and the rest 50% are of size 10. Ordering the text messages by size level $i$, we see that messages of the first level (size 5, $i = 5$) account for 50% of the total messages, and were contributed by 66% of the total user population (2 out of the 3 users). The next level (messages of size 10, $i = 10$) accounts for the other 50%, and was contributed by 33% of the user population. A score $s_i$ can then be calculated per message size level $i$ as follows: $s_i = f_i^* (f_u + 2*f_{u + Au})$, where $f_i$ is the fraction of the messages of size $i$ over the total number of messages, $f_u$ is the fraction of user population contributing messages of size level $i$, and $f_{u + Au}$ is the percentage of the user population contributing more messages than size level $i$. For the above example, $s_5 = 0.5^*(0.67 + 2*0.33) = 0.67$ and $s_{10} = 0.16$. Then the Gini coefficient is calculated as $G = 1 - \sum s_i$. In the above example, $G = 0.17$, which indicates a fairly balanced collaboration in terms of the number of messages contributed. In the case of one user contributing far more messages than the others, Gini increases (e.g., $G = 0.37$ for $u_1 = 5, u_2 = 10$ and $u_3 = 30$, and $G = 0.65$ for $u_1 = 1, u_2 = 1$ and $u_3 = 1000$). We measure the gini coefficient per condition and on the basis of the (a) number of text entries contributed by the participants in the Etherpad collaboration space ($G_N$), and (b) length of the participants’ Etherpad entries ($G_L$). In both cases the SOPs condition involves more balanced collaborations ($G_N = 0.43$, and $G_L = 0.24$) compared to the other two conditions, followed by the No-Agency ($G_N = 0.48$, and $G_L = 0.38$) and then the Placebo ($G_N = 0.53$ and $G_L = 0.42$) condition.

The above results, combined with the fact that SOPs members are more satisfied with their collaboration and have more common working styles than the benchmark condition groups, indicates that overall SOPs seem to collaborate more harmoniously in writing their common stories.

Looking deeper into the mechanics of self-organization

After examining how self-organization affects the work and collaboration quality of participating pairs across the conditions, we now take a deeper look into how this method affects more subtle behavioral elements of collaboration formation.

Strategic voting. So far, we have seen that SOPs members tend to collaborate better, feel more satisfied by their collaboration, and produce better work results. We now look into what motivates people to form the work pairs the way they do. For this, we compare the SOPs and Placebo conditions, since these are the only ones where people were given the option to select their collaborator (although this option is not honored in the Placebo condition). Since the participants of the Placebo condition do not eventually change their work group, we look at voting intention, as revealed during the collaborator selection stage of each condition. In this stage, which takes place after each collaboration session as explained in detail in the “Collaborator selection” section, participants get to indicate which collaborator they would like to work with in the next round. We first examine whether people tend to select previous winners as their preferred collaborators. Indeed, an Analysis of Variance shows that, regardless of condition, the winners of previous rounds gather significantly more profile votes on average ($m_{win} = 5.89, SE = 0.377$) compared to individuals who have never won in the past ($m_{non-win} = 4.04, SE = 0.377$), with $F(1, 98) = 11.431, p = 0.001, \eta^2 = 0.171$.

We also examine whether strategic voting took place in the story voting phase in the form of contender voting, that is, people deliberately downvoting the strongest pairs, that is, previous winners, to afford their own pair a higher chance of winning. Using data from all three conditions (since the voting phase was accessible to all participants) we find that persistent winning pairs (i.e., worker pairs that won a round and stayed together for at least one more time) were not significantly less likely to be downvoted in a subsequent round than one-time winners (those pairs that won previously and did not win in a subsequent round) (39% vs. 61% respectively). A chi-goodness of fit test confirms the above by failing to reject the null hypothesis of equal percentages, with $\chi^2(1, 28) = 1.286, p > .05$. The above results indicate that strategic voting occurred in the collaborator selection phase where people tried to pair with previous winners, but not in the story voting phase.

Winners choose collaborators based on winning potential, non-winners choose those whose profile they like the most.

The aforementioned strategic voting strategy seems to pay off. In the question “What mattered the most when choosing a collaborator?” at the end of the task, we observe that winning participants were twice as likely to report that they chose the person that would make them win (28% of winners, i.e., 22 out of 79) compared to non-winning participants (8 out of 55, i.e., 15% of non-winners). In contrast to winners, who seem to choose strategically their collaborator, non-winners were twice as likely to choose people whose profile information they liked the most (11% of non-winners versus only 4% of winners).

These two elements, that is, winners driven more by a “playing to win” strategy and non-winners driven more by their collaborator’s profile information, were the only answer items that distinguished winners and non-winners, as the rest of the participants’ reported answers to this question received relatively equal percentages of answers (Figure 9). We also note that from the other collaborator selection strategies, choosing based on skill, that is, a prospective collaborator’s individual writing sample, was the one preferred by most participants (almost 50% both in winners and non-winners
alike), but as we saw it did not, in the end, make a real difference as to winning probability.

**Pairs that change often are less likely to win.** Of the total 48 people who were part of the SOPs condition, 46% (22 users) stayed with the same collaborator across all three rounds of the experiment, 42% (20 users) stayed with the same collaborator for two rounds, and the rest 12.5% (6 users) changed collaborator in every round. At the worker pair level of analysis, the SOPs participants created 39 distinct worker pairs (out of the total 72 possible ones, since some pairs chose to stay together for two or even three rounds). 28% of these worker pairs (11 out of 39) stayed together across 3 rounds, 26% (10 out of 39) stayed together for two rounds and the remaining 46% (18 out of 39) of the pairs worked together only once across the three rounds. The worker pairs that stayed together for all three rounds won 55% of the times (10 victories across the total 18 rounds of the SOPs experiments), those that stayed together two times won 28% of the times (5 victories across the total 18 rounds) and those that stayed together only one round won 17% of the times (3 victories across the total 18 rounds). This benefit does not just come from working together more, as that case the benchmark conditions (where each user stays with the same teammate across all rounds) would surpass in quality the SOPs condition. Rather, this result indicates that the exploitation strategy may pay more off than exploration, once the workers have found a suitable collaborator for themselves.

**Network analysis:** SOP participants tend to form more distinct clusters of compatible collaborators. Finally, in a parallelism with machine learning and data-driven self-organization, we look for emergent patterns in the way that people “cluster” across the three rounds, in the two conditions of SOPs and Placebo (the option to self-organization does not apply in the No-Agency condition by default). To do so, we represent the experimental batches as bidirectional affinity graphs, with each affinity graph consisting of a set of users (nodes) and a set of dyadic ties (edges) among user pairs, as explained in the “Self-organization algorithm” section. Each edge between a pair of user nodes receives a value, which corresponds to their “pairwise affinity,” that is, the intent of these two people to collaborate with or avoid one another, as denoted by their voting preferences at the end of each collaboration round. We then apply social network analysis, which provides a set of methods for observing the emerging patterns of the collaborations.

In our analysis of the graphs, we take into account the process of change produced by the dynamic connectivity of the vertices, with every worker accumulating votes across the rounds. The set of \( V \) of vertices of the graph is fixed, whereas the set of \( E \) edges changes with time, in an incremental fashion. The calculation of the final graphs is the same as the sequential analysis technique used for cumulative sum. The networks formed by the partial sums of weights describe the overall interactions between actors across the phases of the experiments. Where a pair of users consistently voted to remain together, the ties between the two nodes shorten, indicating stronger attraction for collaboration. Greater distances between nodes can be formed by repulsion as either one or both the collaborators downvoted the other. The results from a preliminary analysis of the network topology show that the pairs that worked under the SOPs condition create on average more clusters or chains, unlike the benchmark condition, which display for the great majority of the graphs, isolated dyadic clustering. Both in-degree and out-degree weights are considered. Stronger ties between workers creating larger clusters mean the potential creation of channels for interpersonal communication, determined by the person’s choice of one or multiple collaborator/s. More so, social networks formed under the SOPs condition display stronger polarized attraction-repulsion mechanisms that are less detectable in the alternative condition.

Considering together, the above indicate a pattern in human collaboration behavior that is similar to machine learning-based self-organization: when given the agency to form their own virtual collaborations, and an “objective function” to maximize (which in this case is winning a reward), people do tend to actively explore their candidate collaborator space and to gradually form clusters of “compatible” sub-groups that are fairly distinct and separate one from the other. As we also explain in the Discussion section, further studies can be made in this very interesting direction, to examine whether other well-observed in machine learning can also be observed in online work settings, such as the gradual stabilization of the compatible self-formed pair clusters, and how many exploration rounds this would require.

**Discussion, limitations, and future work**

In this work, we make a first attempt to explore the phenomenon of self-organization for online work. Our results show that enabling participants to choose collaborator, and honoring this choice through an algorithm that aims to maximize intra-group preference, yields results of higher quality and worker pairs that are more satisfied by their collaboration. Below, we discuss a number of points that merit further exploration.

**Objective function of self-organized collaborations.** The collaborative/competitive setting of this study encouraged people to choose strategically whom they would pair up with, showing an explicit preference for previous winners. From a macroscopic point-of-view, the worker collective adapted their objective function (i.e., how they make their teaming decisions) to their environment (reward structure), contrary to machine learning-based self-organization, where the objective function is
determined a priori by the algorithm designer. The emergence of the objective function as a collective behavior pattern, is in line with prior work by Woolley and colleagues, who draw attention to a general collective intelligence factor to explain group performance (Woolley et al., 2010, 2015; Engel et al., 2015). In the future, it would be useful to explore how to affect the SOPs objective function in other ways, for instance by removing competitive incentives. This could encourage people to explore more diverse matchings and lead to different collaborative outputs (originality, plot structure, etc.). Another factor that can affect the objective function in a self-organized context is timing. In our study, participants saw the previous round’s winners directly before choosing their collaborators of the next round, which has likely influenced their preferences. Conversely, asking participants to first indicate their preferred collaborators before showing them the winners could have an effect on their choices, and therefore would be an interesting factor to explore as part of future work.

**Exploration-exploitation trade-offs.** During the collaborator selection stage, participants traded between exploration and exploitation. Opting for the former meant that workers increase their chances to find an optimal match, but this comes at the cost of a higher collaboration learning curve. Opting for the latter means that participants can work with an already familiar collaborator, but comes at the cost of missing out on potentially better matches. These trade-offs, which are driven by the behavioral tendencies of the particular batch, fundamentally affect the performance of the collaboration formation algorithm and consequently the performance of the worker collective. Specifically, the participants’ decisions shape the affinity matrix used by the SOPs algorithm to form the worker pairs. When most users opt for exploitation, meaning that they consistently vote to work with the same collaborators, the affinity matrix sparsity increases as less options are available for new pairings. One way to fix this is utilizing network analysis (i.e., Figure 10) to pair participants based on clustering, that is, recommending “collaborators of collaborators.” In the future it would be useful to explore whether such an approach, similar to collaborative filtering used in recommender systems, could help address cold-start problems while still ensuring user agency. An overall risk-averse batch of workers can also make the collaboration formation algorithm susceptible to local optima as people do not explore possibly better collaborators. Future studies could explore introducing randomness in the SOPs algorithm to enable exploration. In recommender systems, this is achieved through serendipity, and future work could introduce it at different degrees and time points of self-organized collaborative systems.

Second, the exploration and exploitation trade-off is affected by task design elements such as number of rounds and batch size. In our study we used three rounds, and most participants reported that this number was sufficient, irrespective of the condition. Nevertheless, we noted that participants were still changing their collaborator selection preferences at the end of the third round. In the future, it would be interesting to examine how many rounds it takes for an average batch to converge, that is, how many rounds it would take before the SOPs collective stops exploring and starts to fully exploiting its stabilized pair formations. Simulations could useful to explore this part of future work, as adding more rounds means lengthier and costlier tasks. Similarly, simulated scenarios could explore how the batch size affects convergence and the performance of the SOPs algorithm.

**Figure 9.** Participants’ self-reported strategies of selecting a collaborator. Winners, that is, people that had been in a winning team at least once, were twice as likely to purposefully select collaborators that would help them win. Non-winners were twice as likely to choose people whose profile information they liked the most. The rest of the teammate choice strategies received similar percentages by both winners and non-winners.
Game-theoretic perspective

As we saw in the previous section, the worker dynamics and—to a certain extent—the task and algorithm designers’ choices can impact the behavior and performance of the self-organized worker collective. In the future, it would be useful to examine this through the lens of cooperative game theory (Chalkiadakis et al., 2012; Branzei et al., 2008). Using this approach, mathematical models of strategic agents representing the online workers can be constructed, and a variety of scenarios can be simulated examining the effects of their interactions and collective tendencies in detail. Possible scenarios include examining the effect of incentives on the workers’ decision to switch collaborators, individual strategic voting scenarios where single workers attempt to retaliate the collaborators that rejected them, or try to increase their own pair’s possibility of winning by not voting high-quality opponent stories, as well as scenarios that predict the formation of strategic coalitions among worker pairs to increase the involved members’ expected profits.

Scaling, timing, and worker cognitive capacity

In our experimental setting, all task stages were time-bounded, which helped meet the requirements of collaboration synchronicity, time management, and attention retention. This technique is closely related to the concepts of timeboxing (Jalote et al., 2004) and iterative development, used in software design and pair programming (Larman and Basili 2003), and known for its benefit in reducing interruptions on focus and flow (Cirillo 2009). Combining time management, complex decision-making, and time-on-task attention depletion is one of the greatest challenges met by our framework. The complexity of this challenge increases with the pool size. For scalability purposes, we rely on concepts such as thin-slicing (Ambady and Rosenthal, 1992) and cognitive heuristics (Metzger and Flanagin 2013), reducing workers’ mental effort and confining the collaboration within fixed (ad-hoc) time slots. Testing other sample stories and topics, as well as different time slots could further the contribution of our research on collaboration formation for creative and complex projects, independently of the task.

Strategic changes in teaming mechanisms. Group dynamics such as changes in motivation and task coordination are part of a set of “synergetic effects” studied for their influences on collaboration and group composition (Larson 2010). While most teams’ early offsets (i.e., first impressions) tend to subsidize over time, short-lived online groups are more affected by short-term impressions and dynamics (Jung 2016), which may include interpersonal frictions (Barsade 2002) and perceived performance (Gully et al., 2002). Whiting et al. (2020)’s study on parallel worlds sees teams as products of elementary interactions molding the collaborators’ internal representations. By experimenting with multiple versions of the same team—hence parallel—Whiting et al. (2020) note the importance of first impressions at shaping team viability, that is, the team’s capacity to sustain collaboration reflecting on members’ willingness to remain (Hackman (1980)). In a similar manner, our SOPs framework offers collaborators first-hand opportunities to “reset” their viability, as users can decide at regular intervals whether they prefer to stay or leave the collaboration.

Competitive self-organization and peer review as built-in quality assurance mechanisms. A typical risk in online work settings is the low level of veracity in worker responses (e.g., concerning collaborator evaluation) as it is often difficult to distinguish good from low-performance workers (Gaikwad et al., 2016). Competitive self-organization on which the SOPs approach is based, is in line with the main idea of the aforementioned paper, in that it also incentivizes workers to select others based on actual performance and compatibility, as this would help them win. As a second built-in quality mechanism,
the SOPs framework uses peer assessment to identify winning stories; a powerful mechanism to enable collective feedback and emerging selectivity (Whiting et al., 2017). Competitive self-organisation and peer assessment complement the importance of agency within the SOPs framework. Relying on agency alone could stifle innovation as workers would not be necessarily interested in seeking the best collaborator in the pool, especially in the presence, for example, of social pressure. By adding the element of competition and peer assessment workers have a stronger motivation to seek others based on ability and performance. In the future, it would be interesting to investigate additional factors such as intrinsic motivators that could encourage truthful responses and high work quality.

**Extending to larger groups.** A long-term scholarly debate exists in small groups’ literature on whether dyads constitute teams, which inspires our following discussion on applying the proposed model to triads or even larger groups. On the one hand, scholars such as Moreland (2010) are of the opinion that the size of a team should be at least three, because dyads are more ephemeral than larger groups, and certain phenomena like majority/minority relations, coalition formation, and group socialization can only be observed in larger groups. Other researchers, such as Williams (2010), argue that two people can be considered to be a team, since some of the most interesting group processes, like inclusion/exclusion, power dynamics, leadership and followership, cohesiveness, social facilitation, and performance occur in dyads in the same manner that they do in larger groups, and that in most instances dyads operate under the same principles that explain group dynamics in teams of three or more. In the experimental part of this study we worked with dyads, as we are primarily interested in the collaborative processes that are linked to worker agency and autonomy, which are already present from the dyad setup (such as inclusion/exclusion, leadership/followership, or performance), and we are less interested in phenomena that occur exclusively in large groups (such as majority/minority relations). This choice was also motivated by the fact that in the particular domain of online collaboration, it is common to work with dyads as the essential foundations for studying collaborative phenomena, which can also have implications for larger groups (Miller et al., 2014; Chikersal et al., 2017; Ahmed et al., 2019; Lykourentzou et al., 2017; Huang and Fu 2013; Rivera et al., 2010). Nevertheless, using dyads is a design choice that limits the application scope of this study to pair work interactions. In the following, we elaborate on how our proposed algorithm and model can be adapted to facilitate larger collaborative structures.

The SOps algorithm first creates a complete graph with candidate group members as nodes and average pairwise ratings as edges, and then produces all possible graph cuts of a given size, which in our experiments was set to two. Next, the algorithm selects those cuts that maximize inter-group affinity, eliminating all alternative groups that the selected individuals could have participated into, until all individuals belong to a work group. The size of the cut is a parameter, and the algorithm can be adapted to compute all possible cuts of a given size, that is, team cardinality (naturally, depending on the specified cut size, the team formed last may have less members). The specific algorithm is greedy, as the problem of optimally partitioning a complete graph into subsets of equal cardinality falls under the NP-Hard complexity category (Feldmann and Foschini 2015). Other approaches, including heuristics and metaheuristics, can also be used to create teams efficiently, for example, the k-nearest neighbor algorithm. In case the number of teams is the goal, polynomial-time algorithms can also be adapted, like the one proposed to solve the Balanced Graph Partitioning problem (Andreev and Racke 2006).

Extending to larger groups may require further adaptations. Currently, the formation of a new work pair is subject to the mutual preference of two users, and the dissolution of an existing pair happens if at least one of the involved users desires to leave the collaboration. With groups of three or more users, adaptations would be required to these rules. Regarding team formation, one can envision two main strategies: team-led, where the majority of existing team members must agree for a new member to join, or worker-led, where a new team is a compilation of either a worker with an existing team or a set of “free workers” based on maximal preference similarity. Similarly, the dissolution of a larger team will be subject to an additional design choice, namely whether a team is considered dissolved even after the exit of a single collaborator (like in worker pairs), or whether there will be a lower minimum threshold of member cardinality before the team disbands. In both cases, the formation and dissolution rules can impact both the individual workers’ degree of agency and the formed teams’ stability, and therefore the performance of the collective.

**Generalizing to other tasks and settings.** In this study, we examined a particular task, that is, fictional story writing, and a particular task design, which is based on iterative cycles of collaboration and competition in real-time synchronous interactions. It is worth considering which other tasks and settings the proposed approach could be applied on, and which adaptations this would require. A real-world task where our approach could be directly applied today is online creative hackathons, for example, those dedicated to video game development. Participants in these settings form temporary groups, comprising artists, developers, and marketers, on a project- and personal preference basis. The same participant can pertain to multiple groups simultaneously. Finding the right group(s) to work with is crucial, especially for independent developers, given the jams’ size, and increasingly online and
deterritorialized character. Currently, creative workers undertake impossible workloads of networking, which undermines sustainable game production (Whitson et al., 2021). Given the inherent similarities of this task with our proposed approach (bottom-up, creativity-centric, and team-centric), we believe that the latter can benefit indie game development. Other tasks of the creative industry sector can also directly benefit from our approach. Collective cinematography and performance arts are, for example, well-suited to coalesce crowd synergies to generate inclusive and immersive participatory projects. Another example are localized cultural events, where online participants collaborate to preserve ethnographic diversity and intangible cultural heritage through local (e.g., culinary) knowledge. Aside the creative sector, other sectors and industries can also directly use our approach, especially those already involving participants in complex, open-ended, and creativity-dependent work. One example is citizen science and public political projects engaging crowd participants’ discussions and propositions in decision-making. Another example is innovation generation beyond idea contests, and “wicked” problem solving. Sustainability projects driven by participants’ ideas and collaboration in solutions for reducing waste and energy consumption belong to this category. A final example are industries using computer vision technologies, like the automotive industry, which could make use of the SOPs approach, engaging crowd workers in debates concerning the prediction, accuracy, and ethics of video surveillance systems, and facilitating human validation where AI still fails at responding to ambiguous inputs. Some of these scenarios could also work without competitive elements, in which case the workers’ satisfaction with their group would be the main drive for optimal collaboration. Naturally, competition could still be supported through incentives of quality, timing, and innovation.

In the future, it would also be useful to explore self-organization for tasks that are more frequently encountered in today’s online work settings, such as those requiring stricter workflows (for example, coordinating programming tasks (LaToza et al., 2015)), those with predefined team roles, tasks requiring expert skills (such as project management consultancy), or those that can be decomposed to smaller work units. Self-organization in those contexts could mean extending worker agency to not only choosing collaborators, but also splitting work responsibilities and delegating tasks among collaborators, in line with latest directions on crowd work reported in the literature (Wood et al., 2019b).

In our setting, workers interacted with one another in real-time. Real-time crowdsourcing has gained research interest in several domains, including audio transcription to support deaf and hard-of-hearing people on-the-go (Lasecki et al., 2012), GUI control (Lasecki et al., 2011), drawing assistance (Limpaecher et al., 2013), but also in collaborative settings where workers continuously correct their task output alleviating the client from the need to repackage results or generate tight feedback loops (Retelny et al., 2014). Real-time and synchronous interactions offer the possibility for quicker results, which is important when the creative task is time-bounded or when all participants are available at the same time, for example, in creative hackathons as we have already discussed. Adapting our approach to an asynchronous setting could help extend it to creative tasks that benefit from longer reflection periods. This adaptation would involve increasing the round duration to longer periods (days or weeks instead of minutes), and enabling an “open call for collaboration” structure where workers without a pair could find one another. Further, our approach is based on distinct iterative rounds of gradual task work and review, until task completion. Extending our approach to tasks or settings without the iterative work property, would require adaptations. One could envision a continuous self-organizing lifecycle, where workers can change collaborators at any given moment. Such an approach could improve the potential of self-organization in the sense that non-functional collaborations would dissolve faster, but it would also mean that a part of the workforce risks becoming temporarily inactive since certain workers may find themselves without work in case no other collaborator is available when they leave their existing group. It would therefore be useful to calculate the benefits over the costs (budget, performance, and worker well-being) of such an adaptation, and consider using methods such as explanations to support workers’ decisions in this context.

Finally, the proposed approach could find application in settings other than work, for example, in an educational context, where user agency (Jahanbakhsh et al., 2017; Bacon et al., 1999) as well as pair collaborations (Miller et al., 2014) have already proven useful. Learning could be a stronger objective in these settings than performance, and this could mean the necessity for certain changes in the task design compared to the design we used in the present study, shifting from a collaborative/competitive mode to purely collaborative. In this case, it might be beneficial to motivate learners to form groups with high skill diversity (expert–non-expert, mentor–trainee, etc.) to promote learning and ensure that no group or learner is left behind, or encourage collaboration formation based on the learning outcomes. Similarly, more diversified metrics could measure success, rather than best group quality. Such metrics could include average quality across all groups, but also group viability, risk of fracture, or sense of psychological safety.

**Theoretical and design implications for future system design.** A first implication concerns the desired level of...
algorithmic involvement in online collaborative systems. This is related to broader directions in mixed-initiative interaction (Allen et al., 1999) and collective intelligence (Malone and Bernstein 2015; Malone et al., 2010) research that studies the duality between AI and human interaction. Our work is driven by the need to mitigate the problems brought by the tight algorithmic supervision of current top-down group formation systems, ranging from inefficient collaboration to significant psychological discomfort. As we show in this paper, incorporating agency can improve collaboration and worker well-being. In the long run, giving workers more control can help create systems that empower and offer more opportunities for personal development.

Despite its many advantages, user agency also comes with ethical concerns and potential risks. Delegating collaboration formation fully to human participants means that some workers may be more sought after than others. This can prove beneficial (e.g., a person knowing that they can work better with a certain co-worker), but it may also be detrimental (e.g., exclusion of specific individuals due to demographic factors, similarity, or familiarity (Hinds et al., 2000)). Recent work (Gómez-Zára et al., 2019) shows that providing individuals with full control replicates systematic inequality, excludes people who do not “look like” ideal collaborators, and leads to segregated collaborations. To mitigate possible selection biases, our system can be designed to foster diversity. Inspired by traditional recommender systems (Kunaver and Požríl 2017), the SOPs algorithm could be parameterized to promote candidate collaborator profiles that the user has not seen or selected before based on feature dissimilarity. Alternatively, the system could be parameterized to reward (monetarily, or with more time, for example) workers for collaborating with people outside their “comfort zone.” Another potential risk is the exclusion of low-ranked users, such as those who received low scores in early evaluations, or newcomers (for versions of the system that allow this). To help those users recover and avoid segregation, the SOPs algorithm can be parameterized to include the element of serendipity, or to explicitly reward “mentorship,” that is, work matchings that mix low-ranked and high-ranked users. Finally, recommendation diversity and hierarchical clustering techniques can be useful to ensure system scalability. In our experiments, we used batches of up to 12 people; however, batches with more workers may be available or necessary for a particular task. To help workers efficiently process dozens or hundreds of profiles, one could envision extending the SOPs algorithm with hierarchical clustering recommendations, starting with suggestions of collaborator “types,” and then allowing users to explore clusters of collaborators similar to their preferred type. It is in such situations where algorithmic involvement can benefit the collaboration formation rather than blocking it.

Another concern is platform disintermediation where workers negotiate, collaborate, and transact with one another or with the client outside the platform boundaries. In such a scenario platforms risk a disruption of their services and economic loss (Srnicek 2017). Constrained self-organization, such as the one advocated by our proposed system, where an algorithm assists worker collaborator negotiations, has been conceptually proposed (Jarrahi et al., 2020) as the golden mean between safeguarding worker autonomy and protecting the digital platforms’ viability.

A fourth design implication relates to the form used by the system to grant worker agency, where one can distinguish two strategies: direct negotiations and mediated approaches. Direct agency negotiations happen when the system enables users to discuss with several other users on whether the latter want to collaborate or not, until all groups are formed (Guimera et al., 2005; Zhu et al., 2013). The advantage of this approach is that it allows workers to fully explore whether a collaboration should proceed and why. The disadvantage is that the negotiation process can be lengthy and the number of possible collaborations that can be explored is minimal due to human cognitive limitations and fatigue. Direct negotiations also risk exposing users’ personal opinions (preferences, rejections) directly to their collaborators. In popular person-to-person recommender systems (David and Cambre 2016) the user first states their intent at system level (e.g., dating apps permitting users to swipe left/right) before allowing the newly matched pair to establish unmediated communication. Mediated agency approaches, like the ones explored in Tacadao and Toledo (2015), (Manya et al., 2017) and (Meulbroek et al., 2019), elicit collaborator preferences using methods such as the maximum satisfiability problem (MaxSAT) (Hansen and Jaumard 1990), the comprehensive assessment for team-member effectiveness (CATME) (Barrick et al., 1998), and the Gale and Shapley (1962) algorithm. The advantage of this approach is that it allows workers to explore significantly more candidate collaborators since they only interact with collaborator profiles. The disadvantage is that it grants less time for reflection on a candidate collaborator’s suitability, and relies more on algorithmic involvement. The approach adopted in this paper fits more in the second category, that is, mediated agency, since the SOPs system uses the participants’ explicit stated (and not deduced) preferences about their collaborators, rather than the outcome of their in-between direct negotiations. This design choice is more appropriate for online and crowd work platforms, where the tasks are time-bounded, and there is a need to maintain scalability in the presence of dozens of candidate collaborator profiles. At the same time however, our system allows users to directly “negotiate” with
one another, in the sense that workers can try out different collaborators across the rounds by actually working with them. Future design frameworks could explore how direct negotiations could be better incorporated into large-scale online work systems without negatively impacting task completion. One possible solution could be combining direct and mediated negotiations and maintaining these in a database across tasks, thus enriching worker’s negotiation history over time.

Limitations. A first limitation of this study concerns the scope and degree of worker agency. Although this is one of the first studies exploring self-organization in an online crowd work context, it did restrict worker agency to teammate choice. Once the dyads were formed for a particular round, workers could not contest their placement for the round’s duration. The rest of the workflow settings, like activity timing, round cardinality, or worker batch size, were not up to the workers to decide, and this inevitably limits their self-management potential. Future work could explore adding a negotiation stage after pair formation, and weigh its benefits with the relative increase in task time, task cost, and worker effort. Future work could also explore giving workers agency over the entire workflow design, to better tailor their collective work strategy to the needs of the particular task.

A second limitation concerns the overheads in terms of task time and cost, and worker cognitive effort. Compared to top-down collaboration formation, where workers do not participate or participate only implicitly in the decision of who will work with whom, the proposed self-organized system increases the time people spend on rating their collaborators, reviewing the collective outcome, and selecting their preferred collaborators. These parts of the task, which can be referred to as self-coordination costs and are extraneous to the work performed on the task itself, make up for almost half the task time (four out of the total 8 minutes per round). These extra coordination costs render the SOPs approach more expensive than top-down approaches. This is a phenomenon that is inherent to self-organizing social collaborative systems, such as Wikipedia (Kittur et al., 2009). Adding even more worker agency, for example, full agency across all task workflow stages, could result in even higher costs. Task designers need therefore to carefully weigh the benefits of self-coordination in terms of creative task quality and worker well-being, compared to more traditional approaches that do not grant worker agency. It is likely that the SOPs approach will yield a higher return-on-investment for tasks that are collaboration-intensive, and where the creative workers’ perspective is important due to the ambiguous nature of the final outcome. Such tasks are most likely those of the creative industry sector, for example, video game development, or the generation of new forms of art. Other tasks that rely less on collaboration and open-ended creativity may not benefit from the proposed approach as much, due to the high coordination costs it involves.

Overheads may also refer to the additional cognitive effort workers need to dedicate to learn to work with a new group. Previous research examining team rotation on creative tasks show that membership change brings new perspectives, which is invaluable for open-ended work and increases its quality (Salehi and Bernstein, 2018). Yet, the same research also confirms that team shuffling can break task familiarity. Collaboration shuffling in self-organizing systems is controlled by the worker collective and not by a rotation algorithm. Future research using explanations as in recommender systems (Tintarev and Masthoff, 2007) could be useful to help workers make more informed collaboration formation decisions and assess whether the benefits of a possible collaborator change outweighs its limitations. Further, the experimental setting used in this study functioned with closed batches of workers, that is, a set of people who participated from the beginning until the end of the task. Given the volatility of crowd work environments it may be necessary to examine a fourth condition in the future, in which workers are paired randomly with a different collaborator at each round. This condition could capture the volatility of crowd workers and bring the proposed approach closer to the reality of current crowd work platforms. A fifth limitation is that in case of a tie in the story voting phase, the algorithm picks one story randomly. A design amelioration could be the use of story branches, similar to those used in software version control systems, where users could follow the story continuation they voted for, for a limited (e.g., one) number of rounds until consensus is reached.

Finally, the design of this study combines competition and collaboration; elements that are often seen in real-life crowdsourcing scenarios (e.g., Kaggle competitions). We retained the element of competition and its surrogate products of pair performance and user popularity as they are both secondary outcomes of rating users for their output. Users in our study can also see their peers’ expertise information, similarly to how people almost always become aware of the expertise of other collaborators in real-world team applications. In our setting, this meant that users tried to pair with previous winners in the collaborator selection phase. Nevertheless performance and popularity can also impact collaborator choice beyond the necessities of the task, and into selection biases. Contender voting, where people deliberately avoid voting for the best collective outcome, although not observed in our setting, could be a risk in settings with higher rewards at stake. To mitigate this, future systems could offer the option to reveal user performance, or any other popularity scores, as an opt-in system component.
Conclusion
This paper investigates the effects of a novel online pair formation framework, titled Self-Organizing Pairs (SOPs). SOPs place increased emphasis on user agency over collaboration formation and relies on the collective decisions of online workers to self-organize into effective pairs while being supported—but not guided—by an algorithm. We compared the SOPs framework with two baselines, where individuals are allocated the same collaborator throughout a creative online task, either with the illusion of user agency (placebo condition) or without it (no-agency condition). Our findings indicate that SOPs lead to a higher quality output, as measured by independent evaluators. Furthermore, we carried out a set of quantitative analyses of the workers’ perceptions of the collaboration, which showed that the pairs formed under SOPs are more satisfied during their common work time, and able to collaborate and help each other more. The purpose of this paper is to lay the ground for other more. The purpose of this paper is to lay the ground for models, which are inspired by bottom-up self-organization. It is our hope that this work will inspire more researchers to focus on the research, authorship, and/or publication of this article.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Ioanna Lykourentzou https://orcid.org/0000-0002-4243-4128

Notes
1. https://wearedynamo.fandom.com/wiki/Guidelines_for_Academic_Requests
2. The Base reward can be either a Base pay or a Base score, depending on the type of participant batch (paid or volunteer); see Section “Reward” for details.
3. Creative self-efficacy is defined as individuals’ beliefs in their ability to produce creative ideas. Past work (Shin et al., 2012) has shown its positive relationship with creative outcome.
4. https://etherpad.org/
5. In the question “The game had three rounds. Ideally, how many rounds would you need to decide who is the best collaborator for you?” at the questionnaire at the end of the task, most participants reported a value close to 3, with no statistically significant difference between the two examined conditions.

References
Ahmed F, Chandra NR, Fuge M, et al (2019) Structuring online dyads: explanations improve creativity, chats lead to convergence. In: Proceedings of the 2019 on Creativity and Cognition. ACM, June 23 - 26, 2019, San Diego CA USA, pp. 306–318.
Allen JE, Guinn CI and Horvitz E (1999) Mixed-initiative interaction. IEEE Intelligent Systems and Their Applications 14(5): 14–23.
Altshuller GS (1999) The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity. USA: Technical innovation center, Inc.
Amabile TM (1998) How to Kill Creativity, Volume 87. Boston, MA: Harvard Business School Publishing.
Amabile TM (2018) Creativity in Context: Update to the Social Psychology of Creativity. Germany: Routledge.
Ambady N. and Rosenthal R. (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin 111(2): 256–274. https://doi.org/10.1037/0033-2909.111.2.256.
Ananny M (2016) Toward an ethics of algorithms: convening, observation, probability, and timeliness. Science, Technology, & Human Values 41(1): 93–117. DOI: 10.1177/0162243915606523.
Andreev K and Racke H (2006) Balanced graph partitioning. Theory of Computing Systems 39(6): 929–939.
Andriopoulos C (2001) Determinants of organisational creativity: a literature review. Management decision 39(10): 834–841.
Argote L (1982) Input uncertainty and organizational coordination in hospital emergency units. Administrative Science Quarterly 27(3): 420–434.
Bacon DR, Stewart KA and Silver WS (1999) Lessons from the best and worst student team experiences: how a teacher can make the difference. Journal of Management Education 23(5): 467–488. DOI: 10.1177/102429789902300503.
Barrick MR, Stewart GL, Neubert MJ, et al. (1998) Relating member ability and personality to work-team processes and team effectiveness. Journal of Applied Psychology 83(3): 377.
Barsade SG (2002) The ripple effect: emotional contagion and its influence on group behavior. Administrative Science Quarterly 47(4): 644–675.
Beck K, Beedle M, VanBennekum A, et al. (2001) The Agile Manifesto. URL: https://agilemanifesto.org/.
Becker BE, Houselid MA, Becker BE, et al. (1998) High performance work systems and firm performance: a synthesis of research and managerial implications. In: Ronald Buckley M, Wheeler AR and Halbesleben JRB (eds) Research in Personnel and Human Resource Management. Japan: Emerald Publishing Limited, 53–102.
Boden MA (2004) The Creative Mind: Myths and Mechanisms. England, UK: Routledge.
Bollen D, Knijnenburg BP, Willemse MC, et al (2010) Understanding choice overload in recommender systems. In:
Proceedings of the fourth ACM conference on Recommender systems30. September 26 - 30, Barcelona Spain. ACM, pp. 63–70.
Bouchard TJ Jr and Hare M (1970) Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology* 54(1p1): 51.
Brajkovich LF (2003) Executive commentary. *Academy of Management Perspectives* 17(1): 110–111.
Branzei R, Dimitrov D and Tjis S (2008) Models in Cooperative Game Theory. Volume 556. Heidelberg, Berlin: Springer Science & Business Media.
Brotchie A and Gooding M (1995) A Book of Surrealism Games. London, UK: Redstone Press.
Carless SA and De Paola C (2000) The measurement of cohesion in work teams. *Small Group Research* 31(1): 71–88.
Carmeli A and Schaubroeck J (2007) The influence of leaders’ and other referents’ normative expectations on individual involvement in creative work. *The Leadership Quarterly* 18(1): 35–48.
Chalkiadakis G, Elkind E and Wooldridge M (2012) Cooperative game theory: Basic concepts and computational challenges. *IEEE Intelligent Systems* 27(3): 86–90.
Chan J, Dang S and Dow SP (2016) Improving crowd innovation with expert facilitation. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. 27 February 2016– 2 March 2016, San Francisco California USA, pp. 1223–1235.
Chen G, Gully SM and Eden D (2001) Validation of a new general self-efficacy scale. *Organizational Research Methods* 4(1): 62–83.
Chikersal P, Tomprou M, Kim YJ, et al (2017) Deep structures of collaboration: Physiological correlates of collective intelligence and group satisfaction. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 25 February 2017– 1 March 2017, Portland Oregon USA. pp. 873–888.
Cirillo F (2009) *The Pomodoro Technique*. NC, USA: Lulu. com.
Cordery JL, Morrison D, Wright BM, et al. (2010) The impact of autonomy and task uncertainty on team performance: a longitudinal field study. *Journal of Organizational Behavior* 31(2–3): 240–258.
Costa AC (2003) Work team trust and effectiveness. *Personnel Review* 32(5): 605–622.
Costa AC, Fulner CA and Anderson NR (2018) Trust in work teams: an integrative review, multilevel model, and future directions. *Journal of Organizational Behavior* 39(2): 169–184.
Cunning G (2013) Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. England, UK: Routledge.
David G and Cambre C (2016) Screened intimacies: tender and the swipe logic. *Social Media+ Society* 2(2): 2056305116641976.
De Dreu CK and West MA (2001) Minority dissent and team innovation: the importance of participation in decision making. *Journal of Applied Psychology* 86(6): 1191.
Di Marzo Serugendo G, Fouvía N, Hassas S, et al. (2004) Self-organisation: paradigms and applications. In: Di Marzo Serugendo G, Karageorgos A, Rana OF, et al. (eds.) *Engineering Self-Organising Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-24701-2, 1–19.
Díaz Suárez S (2015) Evaluating Creative Writing: The Criterion behind Short Stories’ Assessment. Master’s Thesis, Universitat Autonoma de Barcelona.
Dobao AF and Blum A (2013) Collaborative writing in pairs and small groups: learners’ attitudes and perceptions. *System* 41(2): 365–378.
Dow S, Fortuna J, Schwartz D, et al. (2011) Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2807–2816.
Edmonds J (1965) Paths, Trees and Flowers”. *Canadian Journal of Mathematics* 17(3): 449–467.
Engel D, Woolley AW, Aggarwal I, et al (2015) Collective intelligence in computer-mediated collaboration emerges in different contexts and cultures. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. April 18 - 23, 2015, Seoul Republic of Korea. ACM. pp. 3769–3778.
Erdal A and Seferoglu SS (2017) Using etherpad for online collaborative writing activities and learners with different language learning strategies. *Eurasian Journal of Applied Linguistics* 3(2): 205–233.
Faraj S, Pachidi S and Sayegh K (2018) Working and organizing in the age of the learning algorithm. *Information and Organization* 28(1): 62–70. DOI: 10.1016/j.infoandorg.2018.02.005. URL: http://www.sciencedirect.com/science/article/pii/S1471772718300277.
Fausings MS, Jeppesen HJ, Jonsson TS, et al. (2013) Moderators of shared leadership: work function and team autonomy. *Team Performance Management: An International Journal* 19(5–6): 244–262.
Feldmann AE and Foschini L (2015) Balanced partitions of trees and applications. *Algorithmica* 71(2): 354–376.
Feng X, Xu H, Wang Y, et al (2019) The social team building optimization algorithm. *Soft Computing* 23(15): 6533–6554.
Gaikwad SNS, Morina D, Ginzberg A, et al. (2016) Boomerang: rethinking the role of the facilitator in collaborative design. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. pp. 625–637.
Gale D and Shapley LS (1962) College admissions and the stability of marriage. *The American Mathematical Monthly* 69(1): 9–15.
Gard G, Lindström K and Dalhner M (2003) Towards a learning organization: the introduction of a client-centered team-based organization in administrative surveying work. *Applied Ergonomics* 34(2): 97–105.
Gilson LL, Mathieu JE, Shalley CE, et al. (2005) Creativity and standardization: complementary or conflicting drivers of team effectiveness? *Academy of Management Journal* 48(3): 521–531.
Gilson LL and Shalley CE (2004) A little creativity goes a long way: an examination of teams’ engagement in creative processes. *Journal of Management* 30(4): 453–470.
Gómez-Zárate D, Parra M, Twyman M, et al (2019a) Who would you like to work with? In: *Proceedings of the 2019 CHI...
Gómez-Zará D, Parpas M, Twyman M, et al. (2019b) Who would you like to work with? In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–15.

Gorla N and Lam YW (2004) Who should work with whom?: building effective software project teams. Communications of the ACM 47(6): 79–82.

Guimera R, Uzzi B, Spiro J, et al. (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722): 697–702.

Hackman JR and Oldham GR (1976) Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16(2): 250–279.

Hansen MH, Morrow JL Jr and Batista JC (2002) The impact of trust on cooperative membership retention, performance, and satisfaction: an exploratory study. The International Food and Agribusiness Management Review 5(1): 41–59.

Haas M and Mortensen M (2016) The secrets of great teamwork. Harvard Business Review 94(6): 70–76.

Hackett JR (1980) Work redesign and motivation. Professional Psychology 11(3): 445.

Hackett JR and Oldham GR (1976) Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16(2): 250–279.

Hansen P and Jaumard B (1990) Algorithms for the maximum satisfiability problem. Computing 44(4): 279–303.

Heylighen F (2008) Complexity and self-organization. Encyclopedia of Library and Information Sciences 3: 1215–1224.

Hinds PJ, Carley KM, Krackhardt D, et al. (2000) Choosing work group members: balancing similarity, competence, and familiarity. Organizational Behavior and Human Decision Processes 81(2): 226–251.

Ho J, Turnkaya T, Aryal S, et al. (2019) Moving beyond p values: data analysis with estimation graphics. Nature Methods 16(7): 565–566.

Hoegl M and Parboteeah P (2006) Autonomy and teamwork in innovative projects. Human Resource Management 45(1): 67–79.

Hollingshead AB and Brandon DP (2003) Potential benefits of communication in transactional memory systems. Human Communication Research 29(4): 607–615.

Huang SW and Fu WT (2013) Don’t hide in the crowd! Increasing social transparency between peer workers improves crowdsourcing outcomes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 27 April 2013- 2 May 2013, Paris France. pp. 621–630.

Irving RW (1985) An efficient algorithm for the “stable roommates” problem. Journal of Algorithms 6(4): 577–595.
Knijnenburg BP, Willemsen MC, Gantner Z, et al. (2012) Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22(4–5): 441–504.

Kunaver M and Požrl T (2017) Diversity in recommender systems—a survey. *Knowledge-Based Systems* 123: 154–162.

Kurtessis JN, Eisenberger R, Ford MT, et al. (2017) Perceived organizational support: a meta-analytic evaluation of organizational support theory. *Journal of Management* 43(6): 1854–1884.

Langfred CW (2000a) The paradox of self-management: individual and group autonomy in work groups. *Journal of Organizational Behavior* 21(5): 563–585.

Langfred CW (2000b) Work-group design and autonomy: a field study of the interaction between task interdependence and group autonomy. *Small Group Research* 31(1): 54–70.

Langfred CW (2005) Autonomy and performance in teams: the multilevel moderating effect of task interdependence. *Journal of Management* 31(4): 513–529.

Larman C and Basili VR (2003) Iterative and incremental developments: a brief history. *Computer* 36(6): 47–56.

Larson JR (2010) In Search of Synergy in Small Group Performance. Hove, UK: Psychology Press.

Lasecki W, Miller C, Sadilek A, et al. (2012) Real-time captioning by groups of non-experts. *Proceedings of the 25th annual ACM symposium on User interface software and technology*. October 7–10, 2012. Cambridge Massachusetts USA: ACM, 23–34.

Lasecki WS, Murray KI, White S, et al (2011) Real-time crowd control of existing interfaces. In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. October 16–19, 2011, Santa Barbara California USA. ACM. pp. 23–32.

LaToza TD, Di Lecce A, Ricci F, et al (2015) Ask the crowd: Scaffolding coordination and knowledge sharing in microtask programming. In: 2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). Oct. 182015 - 22 , Atlanta, GA, USA. IEEE. pp. 23–27.

Lawler EE III and Worley CG (2006) Designing organizations that support theory. In: *Handbook of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 24(8): 979–1001.

Manya F, Negrete S, Roig C, et al (2017) A maxsat-based approach to the team composition problem in a classroom. In: *International Conference on Autonomous Agents and Multiagent Systems*. 8th - 12th May, 2017, São Paulo - Brazil. Springer. pp. 164–173.

Martin K (2019) Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160(4): 835–850.

McPherson M, Smith-Lovin L and Cook JM (2001) Birds of a feather: homophily in social networks. *Annual Review of Sociology* 27(1): 415–444.

Metzger MJ and Flanagan AJ (2013) Credibility and trust of information in online environments: the use of cognitive heuristics. *Journal of Pragmatics* 59: 210–220.

Meulbroek D, Ferguson D, Ohland M, et al (2019) Forming more effective teams using catme teammaker and the gale-shapley algorithm. In: 2019 IEEE Frontiers in Education Conference (FIE), 16-19 October, 2019, Covington, KY, USA. IEEE, pp. 1–5.

Miller RC, Zhang H, Gilbert E, et al (2014) Pair research: matching people for collaboration, learning, and productivity. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. February 15 - 19, 2014, Baltimore Maryland USA. ACM. pp. 1043–1048.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: mapping the debate. *Big Data & Society* 3(2): 2053951716679679.

Moreland RL (2010) Are dyads really groups? *Small Group Research* 41(2): 251–267.

Nisbett RE and Wilson TD (1977) The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* 35(4): 250.
Okhuysen GA and Bechky BA (2009) 10 coordination in organizations: an integrative perspective. *Academy of Management Annals* 3(1): 463–502.

Pratt AC and Jeuffett P (2009) *Creativity, Innovation and the Cultural Economy: Snake Oil for the Twenty-First Century?* England, UK: Routledge.

Rahman H, Roy SB, Thirumuruganathan S, et al. (2019) Optimized group formation for solving collaborative tasks. *The VLDB Journal* 28(1): 1–23. DOI: 10.1007/s00778-018-0516-7.

Rani U and Furrer M (2021) Digital labour platforms and new forms of flexible work in developing countries: algorithmic management of work and workers. *Competition & Change* 25(2): 212–236.

Rasmussen TH and Jeppeisen HJ (2006) Teamwork and associated psychological factors: a review. *Work & Stress* 20(2): 105–128.

Renzulli JS, Owen SV and Callahan CM (1974) Fluency, flexibility, and originality as a function of group size. *The Journal of Creative Behavior* 8(2): 107–113.

Retelny D, Bernstein MS and Valentine MA (2017) No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proc. ACM Hum.-Comput. Interact* 1(CSCW): 89:1–89:23. DOI: 10.1145/3134724. URL: http://doi.acm.org/10.1145/3134724.

Retelny D, Robaszkiewicz S, To A, et al (2014) Expert crowdsourcing with flash teams. In: *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST ’14. October 5 - 8, 2014, Honolulu/New York, NY, USA Hawaii USAACM. : . ISBN 978-1-4503-3069-5, pp. 75–85. DOI: 10.1145/2642918.2647409. URL: http://doi.acm.org/10.1145/2642918.2647409.

Rivera MT, Soderstrom SB and Uzzi B (2010) Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. *Annual Review of Sociology* 36: 91–115.

Rokicki M, Zerr S and Siersdorfer S (2015) Groupsourcing: team competition designs for crowdsourcing. In: *Proceedings of the 24th International Conference on World Wide Web*, WWW ’15. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3, pp. 906–915. DOI: 10.1145/2736277.2734097.

Saa FE, Downey RG and Lahey MA (1980) Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin* 88(2): 413.

Sacks H, Schegloff EA and Jefferson G (1978) A simplest systematics for the organization of turn taking for conversation. In: Schenkin J (ed) *Studies in the Organization of Conversational Interaction*. UK: Elsevier, pp. 7–55.

Salehi N and Bernstein MS (2018) Hive: collective design through network rotation. *Proc. ACM Hum.-Comput. Interact* 2(CSCW): 151:1–151:26. DOI: 10.1145/3274420. URL: http://doi.acm.org/10.1145/3274420.

Salehi N, Irani LC, Bernstein MS, et al (2015) We are dynamo: overcoming stalling and friction in collective action for crowd workers. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. April 18 - 23, 2015, Seoul Republic of Korea. ACM, pp. 1621–1630.

Salehi N, McCabe A, Valentine M, et al. (2017) Huddler: Convening stable and familiar crowd teams despite unpredictable availability. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW ’17. New York, NY, USA: ACM. ISBN 978-1-4503-4335-0, pp. 1700–1713. DOI: 10.1145/2998181.2998300. URL: http://doi.acm.org/10.1145/2998181.2998300.

Schmitz H and Lykourentzou I (2018) Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing* 1(1): 1.

Senergundo GDM, Foukia N, Hassas S, et al (2003) Self-organization: paradigms and applications. In: International Workshop on Engineering Self-Organising Applications. 15 July, 2003, Melbourne, VIC, Australia. Springer, pp. 1–19.

Shin SJ, Kim TY, Lee JY, et al. (2012) Cognitive team diversity and individual team member creativity: a cross-level interaction. *Academy of Management Journal* 55(1): 197–212.

Siangliulue P, Chan J, Dow SP, et al (2016) Ideahound: improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST ’16. October 16 - 19, 2016, Tokyo Japan. New York, NY, USA: ACM. ISBN 978-1-4503-4189-9, pp. 609–624. DOI: 10.1145/2984511.2984578. URL: http://doi.acm.org/10.1145/2984511.2984578.

Sieg JH, Wallin MW and Von Krogh G (2010) Managerial challenges in open innovation: a study of innovation intermediation in the chemical industry. *R&D Management* 40(3): 281–291. DOI: 10.1111/j.1467-9310.2010.00596.x.

Smicke N (2017) *Platform Capitalism*. Hoboken, NJ, USA: John Wiley & Sons.

Tacadao G and Toledo RP (2015) A generic model for the group formation problem using constraint logic programming. In: 2015 IEEE 15th International Conference on Advanced Learning Technologies. July 6 - 9, 2015, Hualien, Taiwan. IEEE, pp. 307–308.

Tenney ER, Turkheimer E and Oltmanns TF (2009) Being liked is more than having a good personality: the role of matching. *Psychological Science* 20(5): 647–652. DOI: 10.1111/j.1467-9280.2009.02322.x.

Thornburg TH (1991) Group size & member diversity influence on creative performance. *The Journal of Creative Behavior* 25(4): 324–333.

Tintarev N and Masthoff J (2007) An A* algorithm for the constrained shortest path problem. In: *International Work Conference on Artificial Intelligence*, ICAI ’07. June 18 - 20, 2007, Krakow, Poland. Springer, pp. 303–312.

Tintarev N and Masthoff J (2007) A survey of explanations in recommender systems. In: *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST ’07. October 15 - 17, 2007, Hong Kong, China. ACM, pp. 7–16.

Tintarev N and Masthoff J (2007) A* algorithm for the constrained shortest path problem. In: *International Work Conference on Artificial Intelligence*, ICAI ’07. June 18 - 20, 2007, Krakow, Poland. Springer, pp. 303–312.
Valentine MA, Retelny D, To A, et al. (2017) Flash organizations: crowdsourcing complex work by structuring crowds as organizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI ’17. May 6 - 11, 2017. Denver Colorado USA. New York, NY, USA: ACM. ISBN 978-1-4503-4655-9. pp. 3523–3537. DOI: 10.1145/3025453.3025811. URL: http://doi.acm.org/10.1145/3025453.3025811.

Van Mierlo H, Rutte CG, Kompier MA, et al. (2005) Self-managing teamwork and psychological well-being: review of a multilevel research domain. Group & Organization Management 30(2): 211–235.

van Mierlo H, Cv R, Vermunt J, et al. (2006) Individual autonomy in work teams: the role of team autonomy, self-efficacy, and social support. European Journal of Work and Organizational Psychology 15(3): 281–299.

Wang C, Lizardo O, Hachen D, et al. (2013) A dyadic reciprocity index for repeated interaction networks. Network Science 1(1): 31–48.

Whiting ME, Gamage D, Gaikwad SNS, et al (2017) Crowd guilds: worker-led reputation and feedback on crowdsourcing platforms. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 25 February 2017–1 March 2017, Portland Oregon USA. ACM, pp. 1902–1913.

Whiting ME, Gao I, Xing M, et al. (2020) Parallel worlds: repeated initializations of the same team to improve team viability. Proceedings of the ACM on Human-Computer Interaction 4(CSCW1): 1–22.

Whitson JR, Simon B and Parker F (2021) The missing producer: rethinking indie cultural production in terms of entrepreneurship, relational labour, and sustainability. European Journal of Cultural Studies 24(2): 606–627.

Williams KD (2010) Dyads can be groups (and often are). Small Group Research 41(2): 268–274.

Woehr DJ and Huffcutt AI (1994) Rater training for performance appraisal: a quantitative review. Journal of Occupational and Organizational Psychology 67(3): 189–205.

Wood AJ, Graham M, Lehdonvirta V, et al. (2019a) Good gig, bad gig: autonomy and algorithmic control in the global gig economy. Work, Employment and Society 33(1): 56–75.

Wood AJ, Graham M, Lehdonvirta V, et al. (2019b) Networked but commodified: the (dis) embeddedness of digital labour in the gig economy. Sociology 53(5): 931–950.

Woolley AW, Aggarwal I and Malone TW (2015) Collective intelligence and group performance. Current Directions in Psychological Science 24(6): 420–424.

Woolley AW, Chabris CF, Pentland A, et al. (2010) Evidence for a collective intelligence factor in the performance of human groups. Science 330(6004): 686–688.

Zhou S, Valentine M and Bernstein MS (2018) In search of the dream team: temporally constrained multi-armed bandits for identifying effective team structures. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI ’18. New York, NY, USA: ACM. ISBN 978-1-4503-5620-6, pp. 108:1–108:13. DOI: 10.1145/3173574.3173682. URL: http://doi.acm.org/10.1145/3173574.3173682.

Zhu M, Huang Y and Contractor NS (2013) Motivations for self- assembling into project teams. Social Networks 35(2): 251–264.

Appendix

External evaluator’s rating correlations

Table 2 illustrates the Pearson correlation results for the story quality ratings by the external evaluators. The story evaluations were significantly correlated at the 0.01 level (2-tailed).

Estimation graphics

As a companion to the statistical data analyses presented in the Results section of the paper, in this section we present a series of analyses and visualizations using the method of estimation graphics (Ho et al., 2019). The method, which focuses on effect sizes and their uncertainty rather than null-hypothesis significance testing, is useful to provide an additional understanding of the gathered data from a quantitative perspective. In the following three subsections, we present the estimation statistics analysis results on the (i) story quality, (ii) collaboration quality, and (iii) common competencies data. Each analysis includes a pairwise comparison among the three conditions of the experiment (SOPs, Placebo, and No-Agency) using a Cumming estimation plot (Cumming 2013), which is suitable for multiple pairwise comparisons between three or more data groups that typically employ analysis of variance (ANOVA) methods. The effect sizes and CIs are reported as: effect size [CI width x lower bound; upper bound]. 5000 bootstrap samples were taken; the confidence interval is bias-corrected and accelerated. The P value(s) reported are the likelihood(s) of observing the effect size(s), if the null hypothesis of zero difference is true. For each permutation P value, 5000 re-shuffles of the control and test labels were performed.

Story quality evaluations

Grammar. The unpaired mean difference between Placebo and SOPs is 1.18 [95.0% CI 0.916, 1.44]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs.1 is 0.89 [95.0% CI 0.599, 1.18]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is 0.291 [95.0% CI –0.00821, 0.567]. The P value of the two-sided permutation t-test is 0.0572.

Interest. The unpaired mean difference between Placebo and SOPs is 1.27 [95.0% CI 1.02, 1.52]. The P value of the
two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 0.162 [95.0% CI 1.32, 1.39]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.343 [95.0% CI −0.628, −0.0724]. The P value of the two-sided permutation t-test is 0.0154.

Table 2. The story quality ratings by the external evaluators were significantly correlated at the 0.01 level (2-tailed). N = 1960.

|                  | Grammar | Interest | Originality | Plot | Overall |
|------------------|---------|----------|-------------|------|---------|
| Grammar          | 1       | .537**   | .448**      | .559** | .589**  |
| Interest         | .537**  | 1        | .617**      | .753** | .809**  |
| Originality      | .448**  | .617**   | 1           | .550** | .601**  |
| Plot             | .559**  | .753**   | .550**      | 1     | .739**  |
| Overall          | .589**  | .809**   | .601**      | .739**| 1       |

**Originality.** The unpaired mean difference between Placebo and SOPs is 1.07 [95.0% CI 0.824, 1.312]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 1.82 [95.0% CI 1.54, 2.09]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.751 [95.0% CI −1.04, −0.472]. The P value of the two-sided permutation t-test is 0.0.

**Plot.** The unpaired mean difference between Placebo and SOPs is 0.917 [95.0% CI 0.672, 1.17]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 1.36 [95.0% CI 1.07, 1.63]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.442 [95.0% CI −0.723, −0.166]. The P value of the two-sided permutation t-test is 0.0022.

**Overall story quality.** The unpaired mean difference between Placebo and SOPs is 1.61 [95.0% CI 1.37, 1.85]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 1.59 [95.0% CI 1.3, 1.88]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is 0.0188 [95.0% CI −0.259, 0.295]. The P value of the two-sided permutation t-test is 0.895.

**Collaboration evaluations**

**Co-worker collaboration quality.** The unpaired mean difference between Placebo and SOPs is 0.548 [95.0% CI 0.284, 0.824]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 0.949 [95.0% CI 0.651, 1.23]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.401 [95.0% CI −0.74, −0.0607]. The P value of the two-sided permutation t-test is 0.0204.

**Co-worker helpfulness.** The unpaired mean difference between Placebo and SOPs is 0.511 [95.0% CI 0.234, 0.777]. The P value of the two-sided permutation t-test is 0.0002. The unpaired mean difference between No-Agency and SOPs is 0.914 [95.0% CI 0.635, 1.17]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.403 [95.0% CI −0.711, −0.0747]. The P value of the two-sided permutation t-test is 0.0144.

**Own helpfulness.** The unpaired mean difference between Placebo and SOPs is 0.33 [95.0% CI 0.0651, 0.594]. The P value of the two-sided permutation t-test is 0.0134. The unpaired mean difference between No-Agency and SOPs is 0.581 [95.0% CI 0.276, 0.871]. The P value of the two-sided permutation t-test is 0.0002. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.251 [95.0% CI −0.562, 0.0711]. The P value of the two-sided permutation t-test is 0.12.

**Co-worker skillfulness.** The unpaired mean difference between Placebo and SOPs is 0.372 [95.0% CI 0.119, 0.634]. The P value of the two-sided permutation t-test is 0.0046. The unpaired mean difference between No-Agency and SOPs is 0.825 [95.0% CI 0.513, 1.11]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.454 [95.0% CI −0.772, −0.128]. The P value of the two-sided permutation t-test is 0.0046.

**Figure 11** illustrates the mean differences and confidence intervals of the story quality pairwise comparisons.

**Competencies in common**

**Personal values.** The unpaired mean difference between Placebo and SOPs is 0.213 [95.0% CI 0.126, 0.299]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between No-Agency and SOPs is 0.185 [95.0% CI 0.0711, 0.278]. The P value of the two-sided permutation t-test is 0.0036. The unpaired mean difference between Placebo.1 and No-Agency.1 is 0.0274 [95.0% CI −0.022, 0.104]. The P value of the two-sided permutation t-test is 0.267.

**Skill similarity.** The unpaired mean difference between Placebo and SOPs is 0.175 [95.0% CI 0.0395, 0.304]. The P
The unpaired mean difference between No-Agency and SOPs.1 is 0.215 [95.0% CI 0.0595, 0.339]. The P value of the two-sided permutation t-test is 0.0076. The unpaired mean difference between Placebo.1 and No-Agency.1 is 0.0399 [95.0% CI 0.101, 0.174]. The P value of the two-sided permutation t-test is 0.457.

**Figure 11.** The mean differences in story quality evaluations through pairwise comparisons of the SOPs, Placebo, and No-Agency conditions are shown in the above Cumming estimation graph. From left to right and top to bottom the figures display the story quality evaluation comparisons across the axes of (a) Grammar, (b) Interest, (c) Originality, (d) Plot, and (e) Overall. The raw data (measured in a ten-point Likert scale) is plotted on the upper axes; each mean difference is plotted on the lower axes as a bootstrap sampling distribution. Mean differences are depicted as dots; 95% confidence intervals are indicated by the ends of the vertical error bars.

**Work strategy.** The unpaired mean difference between Placebo and SOPs is 0.253 [95.0% CI 0.188, 0.320]. The P value of the two-sided permutation t-test is 0.0002. The unpaired mean difference between No-Agency and SOPs.1 is 0.442 [95.0% CI 0.286, 0.566]. The P value of the two-sided permutation t-test is 0.0. The unpaired mean difference between Placebo.1 and No-Agency.1 is −0.189 [95.0% CI −0.321, 0.052]. The P value of the two-sided permutation t-test is 0.139.
The P value of the two-sided permutation t-test is 0.0172.

Commitment. The unpaired mean difference between Placebo and SOPs is $-0.00344$ [95.0% CI $-0.156, 0.137$]. The P value of the two-sided permutation t-test is 0.873. The unpaired mean difference between No-Agency and SOPs.1 is $0.126$ [95.0% CI $0.0308, 0.289$]. The P value of the two-sided permutation t-test is 0.0728. The unpaired mean difference between Placebo.1 and No-Agency.1 is $-0.129$ [95.0% CI $-0.289, 0.0308$]. The P value of the two-sided permutation t-test is 0.0998.

Figure 13 illustrates the mean differences and confidence intervals of the pairwise comparisons of the common competencies data.
Figure 13. The mean differences in the competencies that the work pairs reported having in common through pairwise comparisons of the SOPs, Placebo, and No-Agency conditions are shown in the above Cumming estimation graph. From left to right and top to bottom the figures display the common items across the axes of: (a) Personal values, (b) Skill similarity, (c) Work strategy, and (d) Commitment.

The raw data (measured in a binary scale, 1 = both members of the worker pair reported having the competency in common, 0 = one or both members of the worker pair did not report having the competency in common) is plotted on the upper axes; each mean difference is plotted on the lower axes as a bootstrap sampling distribution. Mean differences are depicted as dots; 95% confidence intervals are indicated by the ends of the vertical error bars.