Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams

Ioanna Lykourentzou[13]
ioanna.lykourentzou@list.lu

Angeliki Antoniou[2]
angelant@uop.gr

Yannick Naudet[1]
yannick.naudet@list.lu

Steven P. Dow[3]
spdow@cs.cmu.edu

[1] Luxembourg Institute of
Science and Technology
Belval, Luxembourg

[2] University of Peloponnese
Tripoli, Greece

[3] Carnegie Mellon University
Pittsburgh, United States

ABSTRACT

When personalities clash, teams operate less effectively. Personality differences affect face-to-face collaboration and may lower trust in virtual teams. For relatively short-lived assignments, like those of online crowdsourcing, personality matching could provide a simple, scalable strategy for effective team formation. However, it is not clear how (or if) personality differences affect teamwork in this novel context where the workforce is more transient and diverse. This study examines how personality compatibility in crowd teams affects performance and individual perceptions. Using the DISC personality test, we composed 14 five-person teams (N=70) with either a harmonious coverage of personalities (balanced) or a surplus of leader-type personalities (imbalanced). Results show that balancing for personality leads to significantly better performance on a collaborative task. Balanced teams exhibited less conflict and their members reported higher levels of satisfaction and acceptance. This work demonstrates a simple personality matching strategy for forming more effective teams in crowdsourcing contexts.

Author Keywords

Crowsourcing; team formation; personality-based balancing

ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors, Human information processing; H.5.3 Group and Organization Interfaces: Collaborative computing, Computer-supported cooperative work

INTRODUCTION

In traditional work environments, the compatibility of individual personalities in a team can significantly affect collaboration [2]. Personality reflects the way people think, communicate, make decisions, handle stress, and manage conflict [5, 71]. When a team meshes well, its members communicate more effectively, reflect a more positive work environment, exhibit stronger levels of commitment and produce better outcomes [23]. When personalities clash, people experience interpersonal tensions and conflict, and resist team development [5, 22]. In online settings, where team members collaborate virtually rather than face-to-face, personality differences may, in fact, amplify barriers to building and maintaining trust [29].

For relatively short-lived assignments, like those common in online crowdsourcing [42], personality profiling and matching could provide a simple and scalable strategy for effective team formation. However, it is not clear how (or if) personality differences affect teamwork in this novel context where the hiring model is considerably more flexible and the workforce is more transient, interchangeable, and diverse.

This paper examines if bringing workers together based on personality can lead to better team satisfaction and outcomes. Seventy workers from CrowdFlower took a 28-item DISC (dominance, inducement, submission, compliance) personality test [49] and performed a creative advertisement design task [10] in five-person teams. Teams either comprised a balance of personalities (Balanced groups) or a surplus of leader type personalities (Imbalanced groups). We found that balanced groups produced statistically more creative outcomes, as rated by an advertising expert and crowd judges, than imbalanced groups. Balanced groups used less negative language in their interactions and reported significantly higher levels of satisfaction and acceptance, than imbalanced groups, which exhibited a range of conflicts. These results provide implications for how to use personality testing to form effective teams in crowdsourcing context.

RELATED WORK

Personality affects teamwork

Personality and its relationship with individual work performance has been studied extensively over the years. Personality significantly affects the way individuals perceive their work environment, interact and perform within it [8, 18]. Similarly for team-based work, the group's personality composition (GPC) can significantly affect outcomes. Halfhill et al.'s comprehensive review of 31 studies in various face-to-face settings explores how GPC expressed as the mean and variance of personality traits like conscientiousness, agreeableness, openness or emotional stability affects teamwork [26]. This meta-analysis indicates that GPC has a strong effect on team performance, as well as on other team aspects, such as group cohesion, conflict and team viability.

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-3592-8/16/02...$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819979

With respect to team formation, Gilley et al. [23] reviewed different team building theories, while considering group composition, team goals, selection criteria for group members, and personality structures and created an integrated theoretical model for building effective teams. A recurring finding is that groups formed with complementary personalities, where each member brings unique attributes to the team, produce better results and collaborate more effectively [23, 54]. Especially in regards to leadership, it has been found that teams with highly homogeneous leadership styles may lead to poor outcomes due to power struggles (on teams with all extroverted personalities) or a leadership void (on teams with only introverted members) [56].

Although most research in this space examines the impact of personality composition on face-to-face teams [50, 53], a growing body of work investigates how personalities affect teams working at a distance with the help of various communication technologies. Such technologies can enable virtual teams to collaborate, assuming they share a common ground and have only loosely coupled work [57]. Furumo et al. [19] and Holton [29] find that the mix of individual personality traits can significantly affect trust among both face-to-face and virtual teams, and that this effect gets amplified for virtual teams who face additional communication barriers.

Given the value of personality testing for team composition in face-to-face and virtual teams, the present study investigates the effect of personality for emergent forms of work, such as crowdsourcing [16], peer production [43], or innovation tournaments [69]. Personality testing could provide a simple and scalable strategy for effective team formation in the context of short-term, large-scale, and increasingly complex crowd work. However, unlike virtual teams that potentially have long-term working relationships and loyalty to an organization [55], crowd teams typically comprise of workers with largely diverse cultural backgrounds, work cultures, knowledge backgrounds, physical work place conditions, and time zones [64]. Crowd workers can also have significantly different agendas, values, beliefs and interpersonal communication styles [65] - elements known to create interpersonal tension and decrease team productivity [21].

In contrast to face-to-face teams (and virtual teams within a large organization), crowd work may not facilitate the same sense of belonging and commitment due to the short-lived nature of tasks and a flexible hiring model that treats workers as replaceable and interchangeable [36]. Team commitment can lead to better decision-making and problem solving, and higher quality team outcomes [20]. Given the differences between crowd work and more traditional work environments, in terms of how technology mediates interaction, as well as workforce diversity and the hiring model, it remains an open question whether personality testing can be useful for forming crowd teams. This paper investigates: *(how) does personality composition affect team outcomes under the unique conditions of crowd work, and how can crowd platforms improve team formation?*

Teams may enable more complex crowd work

Crowdsourcing often involves soliciting small contributions for short tedious tasks that can be parallelized across a large group of online workers [34]. Strategies to improve crowd work quality include using plurality optimization (optimizing the number of workers allocated per independent microtask) [3], worker-to-task allocation optimization [38, 24], personalizing task recommendations [74], using unsupervised and supervised techniques to infer worker quality [35], applying worker filtering based on reputation-based screening [12] or "golden data" [37], providing feedback using peers or experts [11], or refining task quality evaluation after [72] or during [61] task completion . Further, Kazai et al. [39, 40] found that certain personality traits such as openness and conscientiousness correlate with better outcomes for individual crowd work (specifically task accuracy), in which case a simple survey tool could help identify the most promising workers.

Recent advances in crowd work have explored methods for task decomposition [44] and workflows that bring together dyads [1] or flash teams [62] as a means of achieving complex work products. Enabling teams to collaborate on crowdsourcing tasks may open new opportunities to accomplish more complex, challenging, and creative tasks that require the integrated efforts of people with diverse talents and functional expertise (skills and knowledge backgrounds) [66], like knowledge synthesis and creative problem solving [41].

Recent research examines strategies to improve crowd teamwork, by examining the incentives that enhance team efficiency [31], manipulating the team size and elasticity [62, 76] or using hybrid expert-crowd team structures with mediators to coordinate crowd team communication [58]. Prior work has not yet examined the effect of personality compatibility on crowd teams, despite the increased chances of conflict when placing highly heterogeneous workers together [52] on a creative task that requires frequent interactions [60].

Building on this related work, we hypothesize that forming teams based on personality will have a significant effect on their outcomes. In particular, we contrast teams that comprise a balance of personalities (balanced groups) with teams that have a surplus of leader type personalities (imbalanced groups). We present two main research hypotheses:

- H_1. *Quality of final outcome*. Teams with balanced (complementary) personalities will produce higher quality outcomes, compared to teams with imbalanced personalities.
- H_2. *Perceived effectiveness & achievement emotions*. Teams with balanced (complementary) personalities will report more positive work achievement emotions [59] (motivation, sharing confidence etc.), and they will be more satisfied with their end result, compared to teams with imbalanced personalities.

METHODOLOGY

Personality assessment tool

To form teams based on personality we needed an assessment tool, which: i) allows the extraction of individual worker personality ii) provides information on the effect of personality on group work effectiveness, instead of only providing individual assessments and iii) is relatively easy to deploy in an online team setting. Personality assessment tools such as the

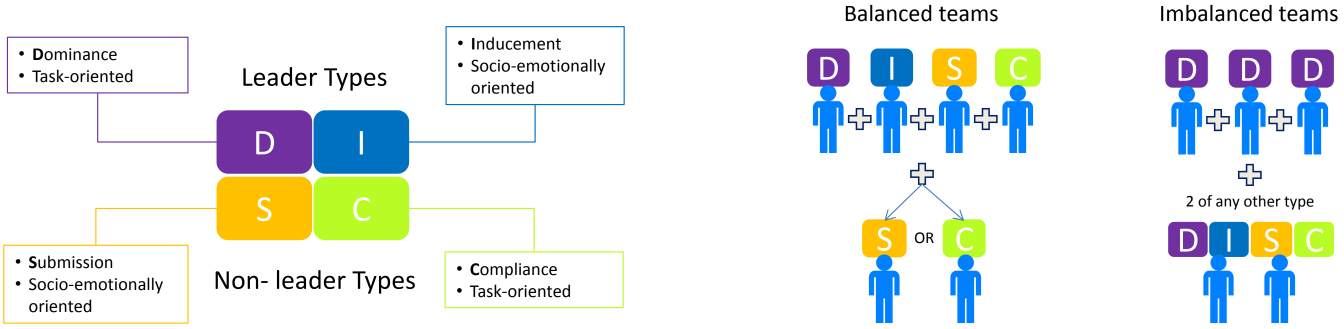

Figure 1. DISC personality test (left) and DISC-based team building. Each team consists of five workers. Balanced teams (middle) have one leader per category (one D type and one I type). Imbalanced teams (right) have at least three D leaders.

Costa and McCrae's [8] NEO-PI-R five factor analysis, Holland's 6 personality types [28], or Eysenck's supertraits [17] primarily study people as *individuals*, with less theoretical development of how to form groups comprising these personality types. As such, these tools fulfill the first but not the second of our study's requirements above. Other works primarily examine the *group* as the unit of analysis, such as the research by Woolley et al. that correlates factors such as gender composition and social perceptiveness with group collective intelligence and group performance in online settings [15, 73]. This line of works focuses on group-level elements (like group collective intelligence), thus fulfilling the second but not the first of our study's requirements. Our work lies in the intersection of the above two approaches, seeking to examine the effect of individual personality as it manifests itself inside a crowd group, affecting and being affected by the roles assumed by the other group members. From the available theories and tests for the assessment of group members (see [23] for a full review), we selected the DISC test, because it explicitly describes how individual personalities interact at the group level and the roles that they will play inside the group. Moreover, the DISC personality test covers four distinct group member types, whereas other tests (like Belbin) require more. Fewer group member roles allows for smaller and more easily manageable groups. Finally, DISC is frequently used as an HR assessment tool in professional evaluation settings for its practicality and tangible outcomes [9, 63], and thus its selection also fulfills the third of our above study's requirements for deployment facility.

DISC draws its principles from temperament theory [71], which examines how personal temperament, i.e. the characteristics of one's personality that remain relatively stable over time, affects the way people behave during one-on-one and group interactions, and which has been studied in detail in the past (e.g. Jung's Archetypes, Myers-Briggs, True Colors, Birkman Method). DISC in particular is designed to focus on the way that different group members will interact with each other and the roles that they will play inside the group. The DISC test (Figure 1) identifies four main types of group members: (i) D-type individuals (leaders, who exhibit high Dominance, are task-oriented and focus on task completion) (ii) I-type individuals (leaders, who exhibit high Inducement, are socio-emotionally oriented and focus on interpersonal relations within the group), (iii) S-type individuals (non-leaders, with high Submission, who are socio-emotionally oriented) and (iv) C-type individuals (non-leaders, who exhibit high Compliance and are task-oriented).

Team formation

Next we defined the team formation elements: team size and personality-based composition. In regards to team size, we needed to take into account two factors: 1) successful teams cover all the group member roles [2], which places a lower bound of four people per group since the DISC tool identifies four distinct member roles within a team (the D, I, S and C roles), 2) there is an upper bound of eight people per team, due to the Ringelmann effect [45], which describes an inverse relationship between the number of people in a group and individual performance and suggests group sizes smaller than eight. From the possible group size values between four and eight, we opted for five people per group taking into account that smaller groups provide administrative flexibility and collaborate more effectively [13], but also not to endanger team coherence in case of a potential worker departure.

In regards to the personality-based team composition, we take into account three parameters: 1) efficient teams cover all group member roles foreseen by the assessment tool [2] (a parameter also used above to determine the lower bound on group size) 2) teams without a leader are not effective [67] and 3) teams with more than two leaders of the same type are not effective either [48]. Taking the above into account, as well as our population demographics (abundance of leader types, as shown later on in this section), we decided to form teams that either balance leadership (while covering all other group roles) or not. Subsequently, the two basic group types that we built and examined were (Figure 1):

- **Balanced groups**, consisting of individuals with compatible personalities. A balanced group consists of: one **D**ominant personality, one **I**nducement personality, one or two **S**ubmission personalities and one or two **C**ompliance personalities. This group type includes all the DISC types but it avoids the presence of two similar types of leaders (e.g. two D or two I types).

- **Imbalanced groups**, consisting of individuals with incompatible personalities. An imbalanced group consists of

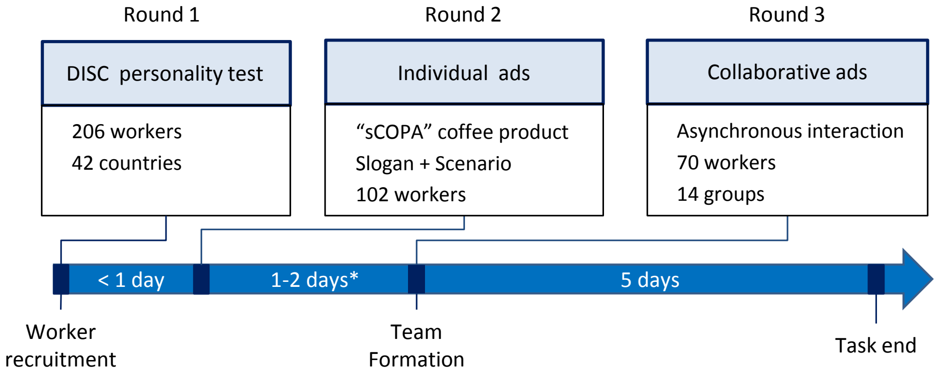

Figure 2. Experiment workflow. (*) Asterisk denotes that the time of worker placement into teams depended on the availability of worker personality types for the team formation.

more than two leaders of the same type (which in our case means multiple D types, due to the abundance of D's in our crowd population).

Task Design

The task we used was *collaborative advertisement creation*. As shown by Dow et al. [10], an advertisement task fulfills key criteria necessary for a crowdsourcing setting: short duration, no requirement for expertise or previous knowledge, ability to express creativity and ability to provide measurements of quality. According to the task, each group of workers was asked to collaboratively create the advertisement campaign of a new product. Generally speaking, a product's campaign can consist of many elements, such as a slogan, scenario, music, logo, broadcasting medium, etc. To keep the task short, asked workers to author the product's slogan (up to 50 words) and scenario (up to 150 words) aimed for TV broadcasting. The product they would have to advertise was a new fictional coffee beverage, called "sCOPA", with the following properties:

> The product is a new coffee beverage from Papua New Guinea. The name of the product is: "sCOPA" standing for COffee PApua new guinea. The product is based on green and brown coffee beans. Green coffee beans are known for their contribution to a healthier diet. The beverage comes in two types: strong and mild. It can be drunk cold or warm. It will be sold both from supermarkets and coffee stores.

Coffee was used, among various candidate products, because it is likely to be known to people across the globe, it has rather neutral connotations (religious, political, etc.), and it has not been exclusively associated with any particular brand (as would be the case for specific soft drink products).

Crowdsourcing workflow design

The overall workflow of the experiment consisted of three rounds (Figure 2).

Round 1. DISC personality test

The first round was an open crowdsourcing task, where workers were invited to take the DISC personality test. As with many personality assessment tools, DISC can be found in longer and shorter versions. We used a relatively short version comprising 28 items, which takes approximately 20 minutes to complete. This task paid $1. Each worker was asked if she would like to participate in the next rounds (subject to selection based on her profile) and, in case of a positive answer, to provide us with a contact email.

Round 2. Individual advertisements

In the second round, the workers who stated interest in participating were invited to make an individual advertisement (slogan and scenario as described above) about the sCOPA coffee product, through a dedicated CrowdFlower job that paid $1. They were instructed that their *"advertisement should be original with a clear market value, using simple, understandable and honest messages and emphasizing the unique aspects of the product."* These instructions were meant to align worker contributions with the final outcome quality axes that we intended to measure at the end of the experiment (see Evaluation metrics section, later on).

Round 3. Collaborative advertisement creation

In the third round, we selected the workers and placed them into groups. Worker selection and assignment to group types was fully random, considering only their DISC personality type and no other information. Moreover worker selection and placement into the groups took into account only those workers that were available at the end of Round 2. This helped avoid a differential selection bias that could affect the results, and prevented any type of personality clashes before the start of Round 3. The participants of each group were given a link to a Google document, on which they would work to create the final sCOPA advertisement. This document contained three parts: 1) task instructions, 2) the five individual advertisements created by the individual team members in Round 2, and 3) a document space to host the final group advertisement. The group members were instructed to read the individual advertisements, and then discuss and create one new advertisement by merging, modifying and taking ideas from any individual advertisement they wanted. Workers were also instructed to actively discuss and interact with the other people in their groups for the final group outcome. The interaction was asynchronous, through threads of comments that the workers could add to the Google document.

Each group had a working period of 5 days. One day before the deadline each worker group was sent a reminder, inviting people to participate if they had not done so. To motivate participation, workers were paid based on their level of interaction with their groups ($0.5-2), while an extra bonus was given to those groups that would manage to make the final advertisement ($1).

Randomness of placement check. The placement of workers into groups took into account only their DISC personality type. No other information, like gender, age, educational background or ethnicity, was used or solicited. However, it was possible to run a post-hoc "diversity" check, using data provided by CrowdFlower regarding the workers' country of origin. This check showed that in 86% of the cases (12 of 14 groups) all group members came from different countries, and in the rest 14% (2 of 14 groups) the country was different for almost all group members (4 out of 5). Thus, worker placement was indeed highly random in respect to a potential influence by cultural proximity. We also checked for potential communication bias, i.e. some teams possibly communicating better due to the higher English proficiency level of its members, compared to other teams. To examine the effects of language proficiency, we tested the distribution of native English speakers across teams. From the 70 workers and 37 different countries that participated in Round 3 (collaborative round), only four workers came from native English speaking countries (one from the USA, two from Great Britain and one from Ireland), and they were evenly distributed across teams and conditions. A posteriori comparison of team performance using one-way ANOVA also showed that there was no statistical significant difference between teams with at least one native English speaking worker and the rest of the teams. Although no other checks are possible, since we opted for asking only the minimum personal data mandated by the hypotheses of the study, this analysis suggests an unlikely influence by other external variables.

Participants

We used the CrowdFlower[1] platform mainly for its breadth of crowd workers, giving access to 5M workers from 154 countries in over 50 labor channels. To ensure a minimum level of English proficiency for all participants, we employed only Performance Level 3 workers (the platform's highest level), who have successfully completed hundreds of different job types, including content writing.

Preliminary population study

Before proceeding with the main part of our study, we ran a preliminary study to understand the sampling of the crowdsourcing population in regards to personality. Overall, from a population of 295 workers who took the personality test of the preliminary phase 56% correspond to leader types (D, I and D/I), 32% to non-leader types (S, C and S/C) and 12% are mixed types (all other combinations) (Figure 3). As can be observed, the crowd worker population is not normally distributed, but the leader types are significantly more than non-leaders. Thus the probability of having a randomly selected team with multiple leaders is high. This observation lends significance to our study: if our hypotheses bear out, this would mean that a team formation that does not take into account personality compatibility in crowdsourcing risks overloading the team with leader types.

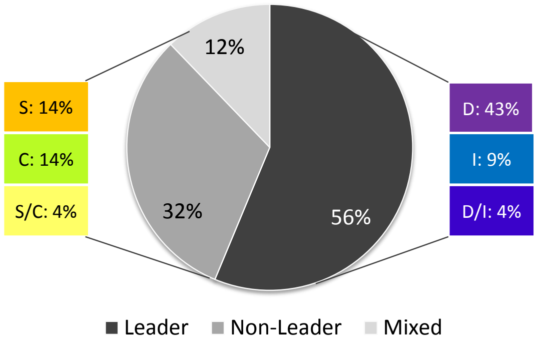

Figure 3. Crowd population personality statistics. The percentage of leader personalities is significantly higher than non-leaders.

Main study participants

After the preliminary study, we proceeded with the main study where the actual manipulation took place. 206 workers (different than the preliminary population study) from 42 different countries participated in Round 1. From these, 102 workers continued to Round 2 and from these 70 people from 37 countries participated in Round 3, which resulted in 14 groups (8 Balanced and 6 Imbalanced). As it can be observed, during the workflow (from Round 1 to Round 3), a significant number of workers dropped out, i.e. did not participate to the next round after being invited. The percentages of personality types for the workers of the main study were very consistent with those of the preliminary study. Ethics approval was obtained and all legal requirements for data protection and information of the participants were fully followed at every step of the process.

Payment selection

Overall, the task paid $2.5-5 depending on participation and performance, calculated based on an average of ∼1 hour of work across the 3 rounds (∼20 minutes on the first round, ∼10 minutes on the second round and ∼30 minutes on the third round). In selecting the payment amount, we took into account three considerations from the literature [14]. First, the payment must conform to the community standards of the crowdsourcing platform used, so as not to bias the end result's quality through workers who would accept a very low payment or workers who would only choose the task purely for its high compensation. Second this payment must be commensurate to the duration of the task and third it must take into consideration the demographics of the target worker population.

CrowdFlower is a micro-payment based platform, similar to AMT, where the median reported crowd worker reservation wage is $1.38/hour [30]. This amount is well below US minimum wage (which is $7.25/hr[2]), but more on balance with the

[1] http://www.crowdflower.com/

[2] We Are Dynamo: Fair Payment, `http://wiki.wearedynamo.org/index.php?title=Fair_payment`

sCOPA: Live life, one bean at a time

A rugged man is free-climbing up Puncak Jaya, the highest mountain in New Guinea. He climbs effortlessly and quickly reaches the summit. Once at the top he takes a moment to admire the view surrounded by clouds and then removes a flask from his backpack, swiftly pouring a brown liquid into the cup of the flask. At this point three brown coffee beans are superimposed at his side with the words "Strong: sCOPA: Live life, one bean at a time". The scene then immediately cuts to the same actor at home relaxing in a bubble bath, drinking from the same flask and at this point three green beans are superimposed at his sides with the words: "Mild: SCOPA: Live life, one bean at a time". Last, the scene cuts to a blank screen with writing saying: "2 beans, 2 sides, which side will you be today?" On either side there is a packet of each type of coffee.

sCOPA - The coffee moment without the moment

A boy and a girl are going to get married without telling their parents, in a register office. The parents from both sides rush to the register office in angry faces. The parents slowly walk to the girl and boy and then the boy and girl give them the sCOPA. The parents drank the coffee and accepted the marriage: "sCOPA changes the moment."

Figure 4. Final advertisement samples. The advertisement of the Balanced group (left) highlights the product's Unique Selling Points (2 types of coffee beans and place of origin-New Guinea) and features detailed scene transition and screenplay. The advertisement of the Imbalanced groups (right) uses stereotypes (age gap, happy couple vs. angry parents), does not highlight the product's unique points and remains at an almost naive screenplay and scene transition level.

average minimum wage across our worldwide population[3]. To ensure fair worker treatment, we monitored worker satisfaction with their payment, during pilots and throughout each phase of the experimental task. At the end of each round, workers (both those that continued and those that eventually dropped out) had the chance to report their satisfaction with their payment through the CrowdFlower "Exit survey." The results of this survey, automatically aggregated by CrowdFlower[4], were 4.3/5 for the first round, 3.9/5 for the second round and 4/5 for the third round, indicating that the chosen payment was considered acceptable by the workers. The fact that workers, including those that later dropped out, were satisfied with their payment, indicates that the selected compensation was appropriate for the specific study setting and it did not affect the workers' decision on whether to participate (or drop out) across the rounds.

Evaluation metrics

Final group outcome evaluation

According to Hoffman [27] a successful ad is creative, honest, plausible, simple (one message is better than two), dramatizes and communicates the reasons to buy the product, rhymes well, and looks for the product's Unique Selling Point (USP - the differences between the product and other similar products). Based on this study, as well as on similar developments on information and content quality research [6], we defined five axes of final group outcome quality: *1) Originality*, *2) Market Value*, *3) Simplicity/Understandability*, *4) Honesty*, and *5) Unique Selling Point (USP)*.

This questionnaire was given to an expert evaluator (advertisement industry professional) and to 700 crowd worker evaluators (different than those participating in the advertisement creation, 50 workers per final advertisement), to get the average user's opinion and to assess the "Wisdom of Crowds" effect (crowds can outperform the estimations of individual experts [68]). Each worker evaluator assessed up to five advertisements, to avoid working memory cognitive overload [51].

Perceived effectiveness & achievement emotions

Following Hypothesis 2, participants after the third round were given a questionnaire developed based on the Achievement Emotions classification study by Pekrun and colleagues [59]. It assessed the participants' perceived emotions regarding their: *1) Motivation*, *2) Comfort Levels*, *3) End Result Quality*, *4) Communication Quality*, *5) Sharing Confidence*, *6) Acceptance*, and *7) Consensus Facility*. The questionnaire also included questions about the participants' Opinion on collaborative tasks and Interest for re-invitation and provided an option for further free-text comments.

RESULTS

Overall, 70 people participated in Round 3, split into 14 groups. Each group discussed and worked together towards the creation of a common advertisement. The average number of comments per group was 22 and the average length of their discussion 500 words. Figure 4 shows advertisement samples created by a Balanced and an Imbalanced team.

Hyp 1: Balanced teams produced better work products

Expert evaluation

The groups' final advertisements were evaluated by an expert advertisement professional for their quality in regards to originality, market value, simplicity, honesty and unique selling point as explained above. The overall score of each advertisement (measured in a scale [0-50], i.e. the sum of scores of the 5 individual axes) was calculated and the scores of the 2 experimental conditions were compared using a Mann-Whitney U test analysis ($n_1 = 6$, $n_2 = 8$, one-tailed).

Mann Whitney was used, due to the small sample size of ratings that necessitated the use of a non-parametric statistical test, which can compare differences between ordinal

[3] `http://www.businessinsider.com/minimum-wage-around-the-world-2015-5`

[4] The task requester can only see overall satisfaction with the task payment and not specific worker answers.

data of two unpaired groups not assuming normal distribution [32]. Balanced groups produced statistically better ($U = 8.5$, $p < .05$) advertisements (mean score $m_B = 26.5$, standard error $SE_B = 5.49$) than Imbalanced groups ($m_I = 10.17$, $SE_I = 3.77$). The same pattern is also observed for each of the 5 individual quality axes (illustrated in Table 1), all of which were statistically confirmed. Figure 5 depicts the overall evaluation rating.

Crowd evaluation

The crowd evaluators ($N = 700$) agreed with the expert regarding the higher quality outcome of the balanced groups (overall mean score $m_B = 29.38$, $SE_B = 0.65$), compared to the imbalanced ones (overall mean score $m_I = 21.48$, $SE_I = 0.99$), as shown in Figure 5. This is statistically confirmed, with one-way ANOVA analysis of $F(1, 12) = 43.99$, $p < .001$. Similar results with p values < .001 were observed for all other individual quality axes. We note nevertheless that the crowd consistently provided higher marks than the expert. The individual performance axes (expert and crowd ratings with their respective mean and standard error values), are presented in Table 1. As it can be observed, according to both the expert and the crowd evaluations, the Balanced groups tended to do better on the axes of Simplicity and Originality, followed by the axes of Honesty and Market Value, and with their least performant axis being that of Unique Selling Point. From the above, hypothesis H_1 is confirmed.

To estimate inter-rater agreement, we used intra-class correlation (ICC), a widely accepted, flexible metric, applicable on group ratings (instead of pairwise ones). Since different reviewer population subsets rated different, randomly assigned advertisement subsets, we use ICC Model 1 - ICC(1) [4]. The analysis gives ICC(1)= 0.73 (with 95% confidence interval [0.583,0.876]), which denote high reviewer agreement on the quality evaluation of the advertisements [47].

Finally, we examined robustness, and more specifically whether there were cases (i.e. in specific groups) that may have a large effect and thus may have distorted the analysis of team performance. We computed Cook's distance influence measures for each observation in our datasets (crowd and expert ratings). All computed influence measures were less than 0.26 for the crowd rating data and 0.21 for the expert rating data, both well within the bounds of a typical threshold of 1.0 [7], and a more conservative threshold of $4/(N - k - 1) = 0.33$ often recommended for small samples [25], where N is the number of observations, and k is the number of parameters. This result suggests that the better performance of the balanced teams (or the worse performance of the imbalanced teams) was not driven by any particular team.

Observations about work products

From a qualitative point-of-view, the advertisements of the Balanced groups are more innovative and touch less conventional topics, in contrast with the advertisements of the Imbalanced groups that seem to revolve around stereotypes (happy couples or families, gender stereotypes etc.). Moreover, the advertisements made by the Balanced teams seem to be more detailed, providing information about how the camera should move (as the advertisements were made for TV broadcasting), how the scenes should transition and so on, in contrast to the Imbalanced teams, who pay less attention to these details. Finally, the Balanced groups seem to highlight more the unique selling points of the advertised product, illustrating for example the contrast between brown and green coffee beans that the product features, the fact that it comes from New Guinea etc. As an example of the above, Figure 4 illustrates one Balanced group advertisement and contrasts it with one Imbalanced group advertisement.

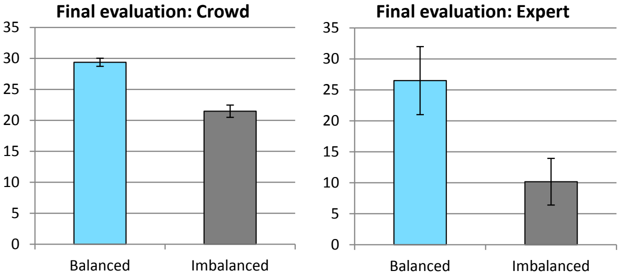

Figure 5. H1 - Final Outcome Evaluation. Balanced groups produced advertisements of higher quality than Imbalanced, as rated by both the crowd (left) and the expert (right).

Hyp 2: Balanced teams reported better communication, acceptance, and satisfaction with end results

To test the statistical significance of Hypothesis 2, we estimated separate linear mixed effects models for each of the dependent measures of the hypothesis (communication quality, acceptance, end result, sharing confidence, consensus facility, motivation, comfort level) to account for the fact that individual participant responses are potentially correlated at group level, due to the communicative experiences among the members of each group. For each of these models, we model the experimental condition (Balanced or Imbalanced) as the fixed effect and group belonging as the random effect (full model). To test for statistical significance, we compare each of these models with information gained from the respective model containing only the random effect (baseline model). Across the dependent measures, when not taking into account the effect of the experimental condition examined by the study (Balanced or Imbalanced) there was indeed considerable dependence of individuals within groups (inter-group correlation), with average ICC across measures = 0.33, indicating that approximately 33% of the total unexplained variability in these measures is explained by between-group differences.

	Expert		Crowd	
	Balanced	Imbalanced	Balanced	Imbalanced
Originality	5.63 (1.12)	2.50 (0.62)	6.02 (0.23)	4.21 (0.29)
Market Value	5.13 (1.22)	1.50 (0.96)	5.83 (0.21)	4.22 (0.27)
Simplicity	5.75 (1.18)	2.33 (0.80)	6.39 (0.25)	5.03 (0.32)
Honesty	5.13 (1.01)	2.00 (0.93)	5.58 (0.24)	4.24 (0.28)
USP	4.88 (1.17)	1.83 (0.79)	5.56 (0.25)	3.78 (0.30)

Table 1. Expert and crowd average ratings per performance axis. Values in parentheses denote standard error. Balanced groups produce consistently better ads across all individual quality axes.

As we will see below, modeling differences between groups in terms of the experimental condition reduces this proportion considerably for three dependent measures (communication quality, acceptance, end result), suggesting that a primary reason that groups were different was because of their experimental manipulation. Overall, three statistically significant results were found (communication quality, acceptance, end result), while the other dependent measures of this hypothesis (sharing confidence, consensus facility, motivation, comfort level) were not statistically confirmed. Figure 6 illustrates these results.

Communication quality higher for balanced groups

The participants of the Balanced groups reported significantly better levels of communication quality ($m_B = 2.58$, $SE_B = 0.1$), compared to the Imbalanced ones ($m_I = 1.9$, $SE_I = 0.09$)), with $B = 0.68$. The p value obtained by likelihood ratio test of the full model (including the effect of condition) against the baseline model (without the effect in question) showed that the full model improves fit significantly over its baseline, with likelihood ratio test $x^2(1) = 15.325$, $p < 0.001$. Additionally, the proportion of total unexplained variability in Communication quality that is accounted for by unexplained between-group variability decreases considerably from $ICC = 0.22$ in the baseline model to $ICC = 3.45e - 16$ in the full model, suggesting groups appear to be different primarily because of their experimental condition. This quantitative result was further validated by qualitative analysis of participant comments. Participants in the Imbalanced teams commented on the lack of activity: *"Unfortunately, I didn't have an active group in which a discussion could be properly held. Either they were a bit inactive, or their communication skills were a bit rusty. They would mostly throw their original idea, and that was about it"* ... *"I would like to have more active members for exchanging more ideas"* ... *"I wish there was more communication."* In contrast, participants in Balanced groups reported more effective communication: *"Everybody collaborated and wrote his opinion, nobody was rude"* ... *"It was fun interacting with other task-ers"* ... *"Felt like I was working with team! I'll be very happy if i could do more tasks like this in future."*

Balanced groups reported higher acceptance levels

Individuals in the balanced condition felt accepted by their groups ($m_B = 2.85$, $SE_B = 0.06$), in contrast to participants of the Imbalanced teams ($m_I = 2.33$, $SE_I = 0.1$) who felt that their group did not sufficiently welcome their contributions. Participation in the Balanced condition affected perceived acceptance levels, increasing it by $B = 0.52$, and with the full model improving fit over its corresponding baseline with likelihood ratio test $x^2(1) = 14.329$, $p < 0.001$. The proportion of total unexplained variability in Acceptance accounted for by unexplained between-group variability decreases considerably from $ICC = 0.32$ in the baseline model to $ICC = 0.057$ in the full model, suggesting groups appear to be different primarily because of their experimental condition. The above is illustrated in this indicative Imbalanced group participant's comment: *"I would have liked a more elaborate slogan but my team didn't approve that"*.

Balanced groups felt more satisfied with end results

People in the balanced groups were more pleased ($m_B = 2.73$, $SE_B = 0.1$) with the group's final result than the people in imbalanced groups ($m_I = 1.8$, $SE_I = 0.13$). Participation in the Balanced condition resulted in an increase of satisfaction with the end result, compared to participation in the Imbalanced condition with $B = 0.93$, $p < 0.001$ and with the full model improving fit over its corresponding baseline with likelihood ratio test $x^2(1) = 18.82$. The proportion of total unexplained variability in End Result satisfaction accounted for by unexplained between-group variability decreases considerably from $ICC = 0.44$ in the baseline model to $ICC = 0.069$ in the full model, suggesting groups appear to be different primarily because of their experimental condition. This finding is also reflected in the qualitative data: *"Despite my encouragement none of the other team members contributed much so the end result was very frustrating really"* (Imbalanced group participant) versus *"We were quite unanimous and satisfied with the end result"*... *"I think that our result was great! It was really a team work so I couldn't ask for something more"* (Balanced group participants).

No differences for other achievement emotions

Finally, there was no statistical difference between the Balanced and Imbalanced conditions with respect to sharing confidence, motivation, comfort level and consensus. Specifically, participants from both group types reported high confidence in sharing their opinion, possibly due to the asynchronous nature of the communication channel. Participants across conditions were also highly motivated to participate, perhaps due to the fact that collaborative tasks are relatively novel in crowdsourcing, making workers more interested in participating. This is further supported by the fact that almost all participants expressed satisfaction with collaborative crowdsourcing tasks and their interest to be re-invited to similar tasks in the future. No statistically significant difference between the groups was found regarding either the comfort levels that the participants had with the process or with their ability to reach consensus. This could be attributed to the asynchronous nature of communication among participants, and the non-sensitive task topic, i.e. coffee advertisement. A

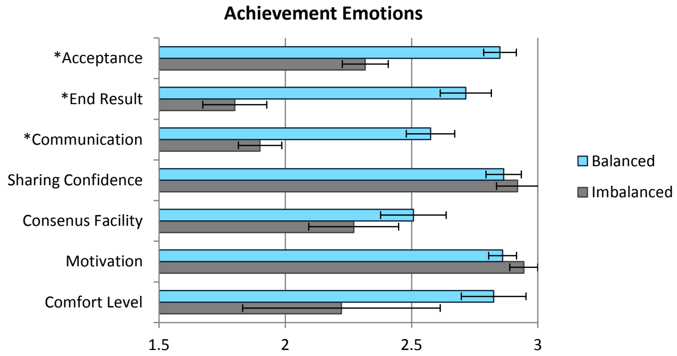

Figure 6. H2. Participant emotions during group interaction and self-perception of group efficiency. Asterisks(*) denote statistically significant axes. Balanced groups reported significantly better communication levels, feeling of acceptance by their group and perceived end result. No significant differences were found regarding the other axes.

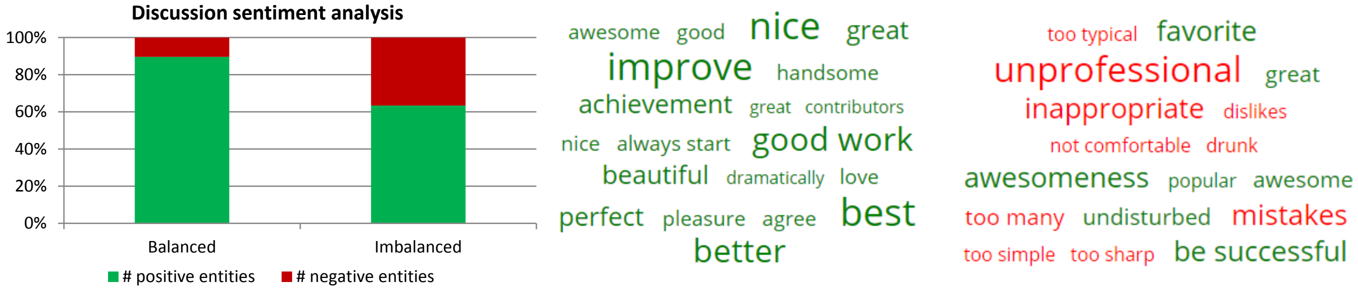

Figure 7. (a) Sentiment analysis (left figure) and indicative Word clouds of team discussions, in Balanced (middle figure) and Imbalanced (right figure) groups. The Balanced groups used more positive vocabulary and expressions compared to the Imbalanced groups. Word/concept frequency indicated by word size. Positive sentiments in green, negative sentiments in red.

different collaborative task, with a more sensitive, polarizing topic (e.g. write an opinion article about terrorism) realized through chat could result in different comfort and consensus levels between the Balanced and Imbalanced groups. From the above, hypothesis H_2 is partially confirmed, for the Communication Quality, Acceptance and End Result axes.

Group discussion analyses

Balanced teams have more positive interactions

Looking deeper into the group discussions revealed several interesting patterns. Tension built up in the Imbalanced groups. Ironic comments were observed *"Are you still in the 90's? You should think out of the box"*, as well as people shouting (using capital letters) to make their point. On the other side, the Balanced groups created a positive and encouraging atmosphere: *"To point it out again: Good job, team!"* . . . *"I hope everything is okay with what we've done, great job everyone and good luck!"* . . . *"It was great working with you all"* . . . *"Yes, good job team! Hope to work with you in the future :-)"*. The sentiment analysis over the group discussions performed using the Semantria software[5] confirms this pattern: the Balanced groups used a significantly higher number of positive vocabulary and expressions while discussing, compared to the Imbalanced groups. Figure 7 illustrates the above, together with a visualization of the content of intra-group communications from one group from each condition.

Observations on how personality types affected the teams

Further interesting observations can be drawn. First, we observed differences between socio-emotional leaders (personality type I) and task leaders (personality type D). The D personalities dominated the group processes in most groups. For instance, D leaders with clear task orientation determined the decision-making processes: *"I placed already all of our slogans below together in the decision page: it will make things easier for us"* . . . *"Let's vote here. Reply with your vote (don't forget own ID). Last one to vote, do us a favor by copy-pasting the winner's (ID, Slogan, Scenario)"* . . . *"OK we have the final slogan. Please now let's build the scenario. Let's work."*. . . *"Hello friends, I leave this comment to remind those who have not yet participated that they have until September 7 to give their [vote, so] that all participate."* Socio-emotional leaders (type I) were focusing on the group interactions, encouraging other members: *"I had a fun time doing it. Congrats to all"* . . . *"That's a great scenario. I like that it emphasizes the concept that the coffee can be drunk either hot or cold. I took the liberty of adding a slogan to the scenario! Feel free to change it if you disagree!"* . . . *"Well I like the first idea since I was the one who wrote it, to be fair it's not that good and it can use some adjustments, tell me what you think, and of all the ideas here I think No 5 is fairly good."*

Second, the behavior of leaders seemed to follow different patterns, depending on the group condition they participated. In Imbalanced groups, the leaders tended to spend their time trying to gain and maintain leadership of the group. They often simply support their original idea, without providing constructive comments. When they do comment on another participant's advertisement, they tend to highlight its negative aspects. For example, below are the comments of one participant who commented negatively on his peers' ads: *"Second scenario: this advertisement is somehow unfinished or like something is missing. Third scenario is imaginary or unreal, I don't like these kind of advertisements, I like real situations and events. Fourth scenario: I don't like it. Too typical, too many people use it as a template"*. This pattern did not allow the groups to make much progress beyond the initial individual ideas, or beyond an initial synthesis of the original ideas, towards an idea maturation stage. This "maturation-ceiling" can also partially explain why the Imbalanced teams produced lower-quality end results.

On the other hand, in Balanced groups the leader personalities acted more towards the interest of the group, discussing with others how to best organize the work or making constructive comments for maturing the initial ideas. A common pattern observed here was that the group leaders initially invited the other participants to comment on each other's advertisements, next they focused the group on 1-2 commonly accepted ideas: *"Well, you're right but still let's make some sense out of it. Let's just stick with the hand bag idea for now since everyone likes it. It's not a big issue"*, and then they helped the group mature their final advertisement, even elaborating on details such as scene filming: *"Yes, great idea. Let's do it like it's from the 80's era, by adding some atmosphere to the story. Thank you for the idea. :-)."* Finally they ensured that

[5] Semantria, Lexalytics. https://semantria.com/

the group will finish on time: *"So we decided then, we are running out of time. Great job everyone and good luck!"*

DISCUSSION

This study examines how personality-based team formation affects crowd teamwork, in terms of the quality of the final outcome and the achievement emotions felt by the workers during their interaction with their team. The results show that placing people together in a way that balances the number of leaders and covers all necessary work roles in the group (as foreseen by the personality assessment tool) significantly improves the end result, the quality of communication, and the workers' perceptions of being accepted and producing good results, compared to teams that have a surplus of leader types.

As to why personality affects crowd teams in such a way, our analysis of group discussions reveals that the imbalance of leadership makes worker teams compete for influence and waste time in struggles regarding who will coordinate the group and whose ideas are going to prevail. We also saw that the imbalance of leadership fosters fixation on the individual members' own ideas and creates negative communication patterns (ironic, tensed atmosphere). Some researchers have identified five stages for effective team development [23, 70], namely: 1) *forming* (team members meet for the first time, accompanied with uncertainty but also enthusiasm), 2) *storming* (members give different ideas and approaches, conflicts occur), 3) *norming* (conflicts lessen and the group finds its work norms), 4) *performing* (roles are clear, members perform their best) and 5) *adjourning* (group wraps-up its result, members part in friendly manner). From these five stages, our study found that the Imbalanced groups only reached the 2$^{\text{nd}}$ stage. On the contrary, most Balanced groups reached the 5$^{\text{th}}$ stage and many of them expressed their wish to work again together as a team in the future.

This study provides practical implications for crowdsourcing task designers, clients and platforms. Through a relatively easy team formation strategy (a short personality test and a subsequent group balancing in terms of personality composition) one can significantly increase team production in collaborative tasks. This strategy can potentially be beneficial for other applications, outside of crowdsourcing, where people work in newly formed groups, such as online learning [46], design challenges (e.g. hackathons, ideation tasks) [33], or corporate team settings.

Our results are important for three main reasons. First, delivering a better team product (like an advertisement) is essential for end-clients, i.e. the stakeholders who pay for the crowdsourcing task. Second, having happy workers is critical for team effectiveness, as pointed out both in organizational psychology [75], and in crowdsourcing [31]. Third, due to the specific demographics of the crowd population that demonstrates an abundance of leader types (as shown in our analysis), crowd teams may be particularly vulnerable to the negative effects of personality incompatibility.

LIMITATIONS AND FUTURE WORK

The extension to other application areas needs to be handled with care: strictly speaking, our results are valid only for the specific work model and population (crowdsourcing) and type of task that was studied (collaborative, creative, of relatively short duration etc.). Other task types could be affected in different ways, or even not at all by personality balancing within the group. For example, there may be less of an effect for routine tasks where personality does not manifest or for tasks where team roles are clearly predefined. Other crowdsourcing populations and other work models might also be differently affected.

Another aspect that could be examined in the future refers to individual performance as a result of personality compatibility and leadership balance inside the group. Specifically, although the present study focused on group performance, and thus gathered and analyzed the respective performance data (i.e. the team advertisements) at the group level, it will be interesting to examine how people's individual performance varies depending on their team's leadership composition, and the resulting personality clashes and compatibilities with the rest of the group members.

Future work could examine these effects under the scope of different crowdsourcing team environments. For example personality can be examined with regards to the mode of interaction. Indeed, in their free-text comments, some participants of our study mentioned that their behaviour would be different if we had used synchronous communication (like chat, or teleconferencing). For example people indicated that had it been a synchronous communication they would be more reserved (*"I would be more quiet. I'm more confident when I'm writing."*) or outgoing (*"If it was a task via chat, I would be more emotional because there wouldn't be time to calm down if I don't like something"*).

Personality could also be examined with regards to task type. For example competitive tasks (like ideation contests among competing crowd teams) may amplify clashes within imbalanced teams, more than collaborative tasks. Furthermore, combining personality with factors like group size, cultural diversity or task difficulty, could potentially lead to more flexible team formation strategies. Future studies could also examine how personality testing affects other populations, like experts (e.g. oDesk-like), volunteers (e.g. open source software development community) or corporate crowds.

CONCLUSION

In this work we examine the impact of personality on crowd team performance. Personality compatibility is an element known to affect traditional face-to-face or virtual teams, but it was unclear if, and how, it affects crowd team work, given the more flexible hiring model and the more transient, interchangeable and dissimilar crowd population. Using the DISC personality assessment tool, we placed crowd workers into two types of teams: Balanced, where all group roles were covered and there was no surplus of leadership, and Imbalanced, which were overloaded with leader personalities. The teams, 14 in total and comprising 5 people each, worked on a collaborative advertisement creation task and their interactions were studied. Results show that the Balanced teams performed better in terms of final outcome quality and reported more positive work-related achievement emotions (perceived communication quality, acceptance by their team and perceived end result) compared to Imbalanced teams. The present study offers a relatively simple team formation strategy, based on personality assessment and matching, which has practical implications for task designers, clients and platforms, as well as potentially for other applications that need to leverage teamwork outcomes.

ACKNOWLEDGMENTS

The authors recognize funding support from the Luxembourg National Research Fund (FNR) under INTER Mobility grant #8734708, and the National Science Foundation under IIS grants #1208382 and #1122206.

REFERENCES

1. Paul André, Robert E. Kraut, and Aniket Kittur. 2014. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, 139–148. DOI: **http://dx.doi.org/10.1145/2556288.2557158**

2. R. Meredith Belbin. 2010. *Management Teams: Why They Succeed or Fail*. Oxford: Butterworth Heinemann, 3rd ed.

3. Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, 33–42. DOI: **http://dx.doi.org/10.1145/2047196.2047201**

4. Paul D. Bliese. 2000. Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*, Steve W. J. Klein, Katherine J.; Kozlowski (Ed.). US: Jossey-Bass, 349–381.

5. Robert Bolton and Dorothy Grover Bolton. 2009. *People styles at work: Making bad relationships good and good relationships better* (second ed.). New York, NY: AMACOM.

6. Kevin Chai, Vidyasagar Potdar, and Tharam Dillon. 2009. Content Quality Assessment Related Frameworks for Social Media. In *Proceedings of the International Conference on Computational Science and Its Applications: Part II (ICCSA '09)*. Springer-Verlag, Berlin, Heidelberg, 791–805. DOI: **http://dx.doi.org/10.1007/978-3-642-02457-3_65**

7. R. Dennis Cook. 1977. Detection of Influential Observations in Linear Regression. *Technometrics (American Statistical Association)* 19, 1 (1977), 15–18.

8. Paul T. Jr. Costa and Robert R. McCrae. 1992. *Revised NEO Personality (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

9. Janina Diekmann and Cornelius J. Knig. (2015). Personality testing in personnel selection: Love it? Leave it? Understand it! In *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice*, J. Oostrom I. Nikolaou (Ed.). Hove, UK: Psychology Press.

10. Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, 2807–2816. DOI: **http://dx.doi.org/10.1145/1978942.1979359**

11. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, 1013–1022. DOI: **http://dx.doi.org/10.1145/2145204.2145355**

12. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2399–2402. DOI: **http://dx.doi.org/10.1145/1753326.1753688**

13. James E. Driskell, Gerald F. Goodwin, Eduardo Salas, and Patrick Gavan O'Shea. 2006. What makes a good team player? Personality and team effectiveness. *Group Dynamics: Theory, Research, and Practice* 10, 4 (2006), 249–271. DOI: **http://dx.doi.org/10.1037/1089-2699.10.4.249**

14. Serge Egelman, Ed H. Chi, and Steven Dow. 2014. Crowdsourcing in HCI Research. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). Springer New York, 267–289.

15. David Engel, Anita Williams Woolley, Lisa X. Jing, Christopher F. Chabris, and Thomas W. Malone. 2014. Reading the Mind in the Eyes or Reading between the Lines? Theory of Mind Predicts Collective Intelligence Equally Well Online and Face-To-Face. *PLoS One* 16;9, 12 (dec 2014), e115212.

16. Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. 2012. Towards an Integrated Crowdsourcing Definition. *Journal of Information Science* 38, 2 (April 2012), 189–200. DOI: **http://dx.doi.org/10.1177/0165551512437638**

17. Hans Jürgen Eysenck. 1975. *The inequality of man.* EdITS Publishers.

18. Adrian Furnham. 1999. *The Psychology of Behaviour at Work: the individual in the organization*. Psychology Press.

19. Kimberly Furumo, Emmeline de Pillis, and David Green. 2009. Personality influences trust differently in virtual and face to face teams. *International Journal of Human Resources Development and Management* 9, 1 (2009), 36–58.

20. James L. Gibson, John M. Ivancevich, James H. Donnelly, and Robert Konopaske. 2009. *Organizations: Behavior, structure, process* (13th ed.). New York: McGraw-Hill.

21. Ann Gilley. 2006. *The manager as change leader*. Westport: Praeger.

22. Jerry W. Gilley and Ann Gilley. 2003. *Strategically integrated HRD: Partnering to maximize organizational performance.* Cambridge, MA: Perseus Books.

23. Jerry W. Gilley, M. Lane Morris, Alina M. Waite, Tabitha Coates, and Abigail Veliquette. 2010. Integrated Theoretical Model for Building Effective Teams. *Advances in Developing Human Resources* 12, 1 (2010), 7–28. DOI: **http://dx.doi.org/10.1177/1523422310365309**

24. Gagan Goel, Afshin Nikzad, and Adish Singla. 2014. Allocating Tasks to Workers with Matching Constraints: Truthful Mechanisms for Crowdsourcing Markets. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 279–280. DOI: **http://dx.doi.org/10.1145/2567948.2577311**

25. Joseph F Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2009. Advanced Diagnostics for Multiple Regression. In *Multivariate Data Analysis (7th Edition)*. Pearson Prentice Hall Publishing.

26. Terry Halfhill, Eric Sundstrom, Jessica Lahner, Wilma Calderone, and Tjai M. Nielsen. 2005. Group Personality Composition and Group Effectiveness: An Integrative Review of Empirical Research. *Small Group Research* 36, 1 (2005), 83–105. DOI: **http://dx.doi.org/10.1177/1046496404268538**

27. Bob Hoffman. 2012. The Ad Contrarian. Fowler Digital Services. ebook. (2012).

28. John Holland. 1973. *Making vocational choices: a theory of careers*. Prentice Hall.

29. Judith A. Holton. 2001. Building trust and collaboration in a virtual team. *Team Performance Management: An International Journal* 7, 3/4 (2001), 36–47. DOI: **http://dx.doi.org/10.1108/13527590110395621**

30. John Joseph Horton and Lydia B. Chilton. 2010. The Labor Economics of Paid Crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce (EC '10)*. ACM, New York, 209–218. DOI: **http://dx.doi.org/10.1145/1807342.1807376**

31. Mokter Hossain. 2012. Users' motivation to participate in online crowdsourcing platforms. In *International Conference on Innovation Management and Technology Research*. 310–315. DOI: **http://dx.doi.org/10.1109/ICIMTR.2012.6236409**

32. David C. Howell. 1992. *Statistical Methods for Psychology*. Belmont: Duxbury Press.

33. Joshua Introne, Robert Laubacher, Gary Olson, and Thomas Malone. 2011. The Climate CoLab: Large scale model-based collaborative planning. In *2011 International Conference on Collaboration Technologies and Systems*. 40–47. DOI: **http://dx.doi.org/10.1109/CTS.2011.5928663**

34. Panagiotis G. Ipeirotis. 2010. Analyzing the Amazon Mechanical Turk Marketplace. *XRDS* 17, 2 (Dec. 2010), 16–21. DOI: **http://dx.doi.org/10.1145/1869086.1869094**

35. Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, 64–67. DOI: **http://dx.doi.org/10.1145/1837885.1837906**

36. Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, 611–620. DOI: **http://dx.doi.org/10.1145/2470654.2470742**

37. Audun Jøsang, Roslan Ismail, and Colin Boyd. 2007. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.* 43, 2 (March 2007), 618–644. DOI: **http://dx.doi.org/10.1016/j.dss.2005.05.019**

38. David R. Karger, Sewoong Oh, and Devavrat Shah. 2011. Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *CoRR* abs/1110.3564 (2011).

39. Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker Types and Personality Traits in Crowdsourcing Relevance Labels. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, 1941–1944. DOI: **http://dx.doi.org/10.1145/2063576.2063860**

40. Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, 2583–2586. DOI: **http://dx.doi.org/10.1145/2396761.2398697**

41. Aniket Kittur. 2010. Crowdsourcing, Collaboration and Creativity. *XRDS* 17, 2 (Dec. 2010), 22–26. DOI: **http://dx.doi.org/10.1145/1869086.1869096**

42. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, 1301–1318. DOI: http://dx.doi.org/10.1145/2441776.2441923

43. Aniket Kittur, Bryan Pendleton, and Robert E. Kraut. 2009. Herding the Cats: The Influence of Groups in Coordinating Peer Production. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*. ACM, New York, NY, USA, Article 7, 9 pages. DOI: http://dx.doi.org/10.1145/1641309.1641321

44. Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, 43–52. DOI: http://dx.doi.org/10.1145/2047196.2047202

45. David A. Kravitz and Barbara Martin. 1986. Ringelmann Rediscovered: The Original Article. *Journal of Personality and Social Psychology* 50, 5 (May 1986), 936–941.

46. Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S. Bernstein, and Scott R. Klemmer. 2015. Talkabout: Making Distance Matter with Small Groups in Massive Classes. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, 1116–1128. DOI: http://dx.doi.org/10.1145/2675133.2675166

47. James M. LeBreton and Jenell L. Senter. 2008. Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods* 11, 4 (2008), 815–852. DOI: http://dx.doi.org/10.1177/1094428106296642

48. Rensis Likert. 1969. *The human organization.* Mcgraw-hill: New york.

49. William Moulton Marston. 1979. *Emotions of Normal People.* Persona Press Inc.

50. John Mathieu, M. Travis Maynard, Tamm Rapp, and Lucy Gilson. 2008. Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. *Journal of Management* 34, 3 (2008), 410–476. DOI:http://dx.doi.org/10.1177/0149206308316061

51. George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 8197. DOI: http://dx.doi.org/10.1037/h0043158

52. Eric Molleman. 2005. Diversity in Demographic Characteristics, Abilities and Personality Traits: Do Faultlines Affect Team Functioning? *Group Decision and Negotiation* 14, 3 (2005), 173–193. DOI: http://dx.doi.org/10.1007/s10726-005-6490-7

53. Richard L. Moreland, John M. Levine, and Melissa L. Wingert. 1996. Creating the ideal group: Composition effects at work. In *Understanding group behavior Vol. 2: Small group processes and interpersonal relations*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 11–35.

54. Paul M. Muchinsky and Carlyn J. Monahan. 1987. What is person-environment congruence? Supplementary versus complementary models of fit. *Journal of Vocational Behavior* 31, 3 (1987), 268 – 277. DOI: http://dx.doi.org/10.1016/0001-8791(87)90043-1

55. Michael Muller, Kate Ehrlich, Tara Matthews, Adam Perer, Inbal Ronen, and Ido Guy. 2012. Diversity Among Enterprise Online Communities: Collaborating, Teaming, and Innovating Through Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, 2815–2824. DOI: http://dx.doi.org/10.1145/2207676.2208685

56. George A. Neuman, Stephen H. Wagner, and Neil D. Christiansen. 1999. The Relationship between Work-Team Personality Composition and the Job Performance of Teams. *Group & Organization Management* 24, 1 (1999), 28–45. DOI: http://dx.doi.org/10.1177/1059601199241003

57. Gary M. Olson and Judith S. Olson. 2000. Distance Matters. *Hum.-Comput. Interact.* 15, 2 (Sept. 2000), 139–178. DOI: http://dx.doi.org/10.1207/S15327051HCI1523_4

58. Cheong Ha Park, KyoungHee Son, Joon Hyub Lee, and Seok-Hyung Bae. 2013. Crowd vs. Crowd: Large-scale Cooperative Design Through Open Team Competition. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, 1275–1284. DOI: http://dx.doi.org/10.1145/2441776.2441920

59. Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P. Perry. 2011. Measuring emotions in students learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology* 36, 1 (2011), 36 – 48. DOI:http://dx.doi.org/10.1016/j.cedpsych.2010.10.002

60. Lisa Hope Pelled, Kathleen M. Eisenhardt, and Katherine R. Xin. 1999. Exploring the Black Box: An Analysis of Work Group Diversity, Conflict, and Performance. *Administrative Science Quarterly* 44, 1 (1999), pp. 1–28. http://www.jstor.org/stable/2667029

61. Aditya Ramesh, Aditya Parameswaran, Hector Garcia-Molina, and Neoklis Polyzotis. 2012. *Identifying Reliable Workers Swiftly*. Technical report.

62. Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S. Bernstein. 2014. Expert Crowdsourcing with Flash Teams. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, 75–85. DOI: **http://dx.doi.org/10.1145/2642918.2647409**

63. James H. Reynierse, Dennis Ackerman, Alexis A. Fink, and John B. Harker. 2000. The Effects of Personality and Management Role on Perceived Values in Business Settings. *International Journal of Value-Based Management* 13, 1 (2000), 1–13. DOI: **http://dx.doi.org/10.1023/A:1007707800997**

64. Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, 2863–2872. DOI: **http://dx.doi.org/10.1145/1753846.1753873**

65. Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 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 (CHI '15)*. ACM, New York, 1621–1630. DOI: **http://dx.doi.org/10.1145/2702123.2702508**

66. Kai Sassenberg, Kai J. Jonas, James Y. Shah, and Paige C. Brazy. 2007. Why some groups just feel better: The regulatory fit of group power. *Journal of Personality and Social Psychology* 2 (2007), 249–267.

67. Leonard R. Sayles and George Straus. 1966. *Human behaviour in organizations.* Englewood Cliffs, New Jersey: Prentice Hall.

68. James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.

69. Christian Terwiesch and Karl Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*. Harvard Business Press.

70. Bruce W. Tuckman and Mary Ann C. Jenson. 1977. Stages of small-group development revisited. *Group & Organization Management* 2 (1977), 419–427.

71. Judy Whichard and Nathalie L. Kees. 2006. *Manager as facilitator*. Hartford, CT: Praeger.

72. Jacob Whitehil, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 2035–2043.

73. Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688.

74. Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2014. TaskRec: A Task Recommendation Framework in Crowdsourcing Systems. *Neural Processing Letters* (2014), 1–16. DOI: **http://dx.doi.org/10.1007/s11063-014-9343-z**

75. Isik Zeytinoglu. 2005. *Satisfied Workers, Retained Workers: Effects of Work and Work Environment on Homecare Workers' Job Satisfaction, Stress, Physical Health, and Retention*. Technical Report RC1-0965-06. Canadian Health Services Research Foundation.

76. Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1445–1455. DOI: **http://dx.doi.org/10.1145/2531602.2531718**