Reviewing versus Doing: Learning and Performance in Crowd Assessment

Haiyi Zhu, Steven P Dow, Robert E Kraut, Aniket Kittur
Human Computer Interaction Institute
Carnegie Mellon University
{haiyiz, spdow, robert.kraut, nkittur}@cs.cmu.edu

ABSTRACT

In modern crowdsourcing markets, requesters face the challenge of training and managing large transient workforces. Requesters can hire peer workers to review others' work, but the value may be marginal, especially if the reviewers lack requisite knowledge. Our research explores if and how workers learn and improve their performance in a task domain by serving as peer reviewers. Further, we investigate whether peer reviewing may be more effective in teams where the reviewers can reach consensus through discussion. An online between-subjects experiment compares the tradeoffs of reviewing versus producing work using three different organization strategies: working individually, working as an interactive team, and aggregating individuals into nominal groups. The results show that workers who review others' work perform better on subsequent tasks than workers who just produce. We also find that interactive reviewer teams outperform individual reviewers on all quality measures. However, aggregating individual reviewers into nominal groups produces better quality assessments than interactive teams, except in task domains where discussion helps overcome individual misconceptions.

Author Keywords

Crowdsourcing, Review, Assessment, Learning.

ACM Classification Keywords

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – Collaborative computing, Computer-supported cooperative work, Web-based interaction; K.4.3 [Computers and Society]: Organizational Impacts – Computer supported collaborative work.

General Terms

Management, Experimentation.

INTRODUCTION

Crowdsourcing has become a powerful paradigm for accomplishing work at a scale and speed previously impossible. Most online crowdsourcing markets currently have a flat organizational structure where requesters post tasks for workers and then decide retrospectively which workers to pay. Recent crowdsourcing research suggests that peer review strategies can help requesters achieve better results because it provides training for the workers receiving assessment [5, 8]. However, peer review strategies incur real costs for requesters, and thus may be marginally valuable, especially if reviewers lack domain knowledge.

Theory from organization science (e.g., [36]) and learning science (e.g., [20]) would suggest that reviewing others' work can offer an opportunity to make a meaningful impact, induce a sense of responsibility, and improve motivation to master a task domain. This paper explores ancillary benefits—such as learning and performance increases—that peer reviewing can provide to the *reviewers*, not just to workers.

Further, while peer review strategies may provide numerous potential benefits, there is relatively little evidence about how to structure it effectively for crowds. An individual reviewer might not have the complete set of skills needed to be an effective reviewer, especially for subjective or generative tasks where gold standards do not apply. Small teams can help overcome individual limitations by bringing together varying perspectives, knowledge, and expertise [18]. However, interactive teams can endure coordination costs [18,32] and are subject to group polarization [9]. *Nominal groups* aggregate independent actions of individuals and provide a cost-equivalent alternative to interactive teams. Prior research shows that nominal groups can overcome production loss and groupthink, especially during generative brainstorming tasks [10, 23, 32]. Thus, this research also investigates whether peer review quality improves by supporting interactive team versus nominal group reviews and whether team size has an effect.

To understand the tradeoffs on learning and performance for peer review versus production work and for individual versus team strategies, we conducted a between-subjects experiment with workers (N=680) from Amazon's Mechanical Turk [12] across five common task domains. In a three-stage procedure (see Figure 1), participants first completed a single production task (pre-task). Then, we randomly assigned participants to a condition where they either completed a review task or a production task, either individually or in teams (main-task). Finally, participants received an

CSCW '14, February 15–19, 2014, Baltimore, Maryland, USA.

email message inviting them to perform more production tasks (post-task). To measure learning, we compared participants' performance across conditions on the pre- and post-task production work. To measure performance on peer review tasks, we calculated the consistency among reviewers, internal coherence, agreement with judges, subjective ratings of the review's helpfulness to production workers, and the review correctness for a mathematics problem.

The study showed that participants who performed peer review tasks had a greater increase in performance between the pre- and post-task than those who only performed production tasks. This indicates that participating in review tasks improves learning in a task domain.

Comparing the different methods for organizing reviewers, we found that interactive reviewer teams outperformed individuals on most quality measures. Moreover, nominal groups outperformed interactive reviewer teams, except when judging the correctness of a mathematics problem. These performance differences were not affected by the size of the team. Also, the organizational structure (individual vs. team vs. nominal) did not affect the performance difference between the pre and post task.

THEORY AND HYPOTHESES

Crowdsourcing researchers and practitioners have recently become interested in incorporating management, feedback and assessment mechanisms in micro-task market places. For example, Mobileworks introduced management strategies to micro-task markets [16]. However, there remains a need for a rigorous scientific foundation for understanding the effectiveness of feedback and assessment mechanisms in crowd work. In the following section, we build on relevant theory to predict the effects of participating in review tasks and the effectiveness of different ways of organizing reviewers in crowd work.

Learning by Reviewing

Reviewing, which is the action of providing information regarding one's task performance (sometimes referred to as providing feedback or assessment), has been investigated in depth in organization and learning science due to its important role in management and education, respectively. Research on organizational behavior shows numerous motivational and performance-oriented benefits of encouraging peers to participate in managerial activities such as evaluating performance and providing feedback, variously termed as self-leadership [21], distributed leadership [1], empowered leadership [29], and shared leadership [26, 37, 39]. Exercising managerial influence such as evaluating others can improve workers' performance by giving them an opportunity to express their voice and make a more meaningful impact [26]. This benefit is consistent with Ryan and Deci's self-determination theory, which argues that people gravitate towards work that allows them to feel competent and autonomous [27].

In the learning sciences, we see a parallel line of research about learning by mentoring [4, 20]. Mentoring also involves providing assessment and instructions to others. Chase and her colleagues demonstrated the so-called "protégé effect" where students made greater effort and learned more when they mentored others [4]. Mentoring others invokes a sense of responsibility that motivates learning, and protects students' egos when confronted with failure [4]. Providing students with opportunities to mentor others can also help them learn by developing meta-cognitive knowledge and skills, that is, students "learn how to learn" [3, 34].

There has been some recent research on peer feedback and reviewing in crowd work. Horton [8] investigated the effects of peer feedback on feedback providers' productivity, finding that evaluating high-output work raised the feedback provider's subsequent productivity compared to evaluating low-output work. Our work similarly focuses on the role of feedback providers, but explicitly compares feedback providers and production workers and additionally looks at learning outcomes. Dow et al. [5] investigated the effects of "self-assessment" (a hybrid of assessing and producing) on learning, showing that self-assessment helps workers improve their task performance over time. However, unlike our study this feedback was self-provided and furthermore did not disentangle assessment from production work. Therefore, our first hypothesis is as follows.

> *H1. Participating in review tasks enhances performance in the task domain.*

Strategies for Organizing Peer Reviewers

While promoting workers to reviewer roles can help them become better workers, they need a considerable set of skills to be an effective reviewer [36]. Reviewers need technical skills (i.e., knowledge about methods, processes, procedures and techniques for conducting a specialized activity) and communication skills (i.e., ability to communicate the judgment clearly and effectively). A single worker might not have the complete skill set to effectively perform review tasks. In contrast, a group of workers acting together are more likely to collectively have the skills needed to effectively review other workers' work [17].

Assuming that groups outperform individuals, requesters still need effective methods to organize people from different backgrounds and perspectives to accomplish a common task. Two alternative methods are plausible: interactive groups and nominal groups [10, 18, 23, 25]. In interactive groups, people interact and communicate with each other while working, while in nominal groups people work individually and their efforts are combined algorithmically, as an average or sum.

Interactive teams might perform worse than nominal teams in a review task because social influence occurring in active discussions might cause the group judgment to move away from the "truth" and towards each other's initial biased positions. The phenomenon is referred to "group polarization" in literature [9]. For example, Schkade et al. reported a study of over 500 mock juries composed of over 3000 jury eligible citizens. They found that deliberation produces a "severity shift" in which the jury's dollar verdict is systematically higher than that of the median of its jurors' pre-deliberation individual judgment [28]. In addition to avoiding unhealthy social influence, nominal teams also avoid other factors that inhibit productivity, such as 1) production blocking preventing members from simultaneously contributing during discussions [30]; 2) free riders do not contribute [10]; 3) time and effort diverted from production to coordinate and resolve conflicts [32]; and 4) "bad apples" who discourage others [6].

Furthermore, although the resources increase with number or people in a group, coordination and other process losses also increase [23]. Therefore, we predict that the performance gap between nominal and interactive teams will increase with group size. To summarize, we hypothesize that:

H2a. Interactive reviewer teams perform better than individual reviewers.

H2b. Nominal reviewer teams will perform better than interactive reviewer teams.

H2c. The performance gap between nominal and interactive teams will increase with team size.

METHOD

Experiment Design

We designed a three-stage experiment: pre-task, main experimental task, and post-task. In the main task, we examined two types of work (production vs. review) and five ways of grouping people (individually vs. four different-sized teams), giving us ten independent conditions (see Figure 1 for details). Half of the participants were in "producer" conditions where they completed five jobs: a writing task, a brainstorming task, a moral judgment task, a mathematics problem, and a summary task. The other half of participants were in "reviewer" conditions where they were

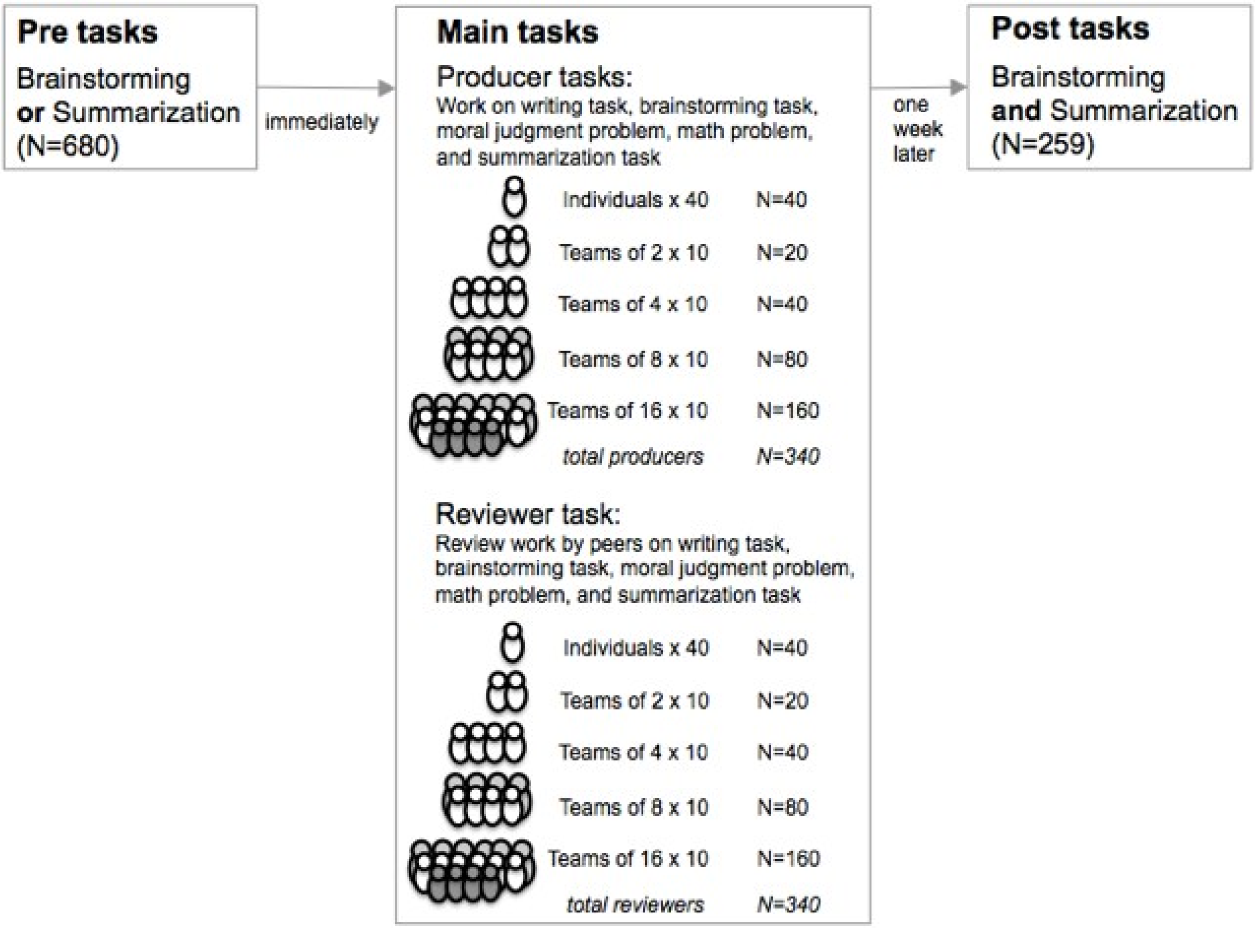


Figure 1: We conducted a three-stage between-subjects experiment with workers from Amazon's Mechanical Turk. To measure learning, we compared participants' performance across conditions on the pre- and post-tasks. *Note that we designed post-task as a separate task to ensure that the baseline study (pre-task plus main task) was not too time-consuming, and thus limit the number of study dropouts. For the same reason, participants were only required to complete one task (either brainstorming or summarization) in the pre-task. Our analysis compared the participants' pre- and post-task difference for only the participants who completed all the three stages.*

required to assess others' work on those same five tasks. For both types of tasks, participants either worked individually or in interactive teams with an exponentially increasing number of workers—two, four, eight or sixteen. In the pre-task, half the participants completed a summary task, and the other half completed a brainstorming task. Then participants were randomly assigned to an experimental condition. After the workers finished the main experimental task, we sent emails to invite participants to complete a post-task, which included both a brainstorming task and a summary task.

Production Tasks

We designed tasks for the "producer" conditions according to McGrath [22]'s taxonomy of group tasks, which cover a variety of different skills and group processes.

- *Writing task*. Participants were asked to write a consumer review for a product from the following list: MP3 player, Mobile phone, Digital camera, Headphones, Game console, Computer, and Laptop. Dow et al. previously used this experiment task to study feedback for crowd workers [5].
- *Brainstorming task.* Participants were asked to generate ten ideas for alternative uses for a brick or a pencil [35]. The more creative, the better.
- *Moral judgment*. Participants were asked to judge a moral dilemma [35]. In this case, a college teaching assistant accepted a bribe from a star basketball player to change his exam grade. The workers were required to make decisions that satisfy the conflicting interests between the basketball team (who wanted to retain the player) and the course instructor (who wanted to punish the cheating behavior).
- *Mathematics problem*. Participants were asked to solve a mathematics problem: "A bag contains two red balls, three blue balls and five green balls. Balls are drawn at random. How many balls do you need to draw so that at least two of them are of the same color?"
- *Summarization.* Participants were asked to summarize a paragraph of text in two or three sentences.

Review Tasks

For the review task, participants were asked to assess five production tasks (one of each type). Participants first made Yes/No judgments about the quality of different dimensions of a worker's answer. For example, to assess the summarization task, participants judged whether the summary was accurate or not. Details of the review tasks are shown below. Then participants decided whether they wanted to approve the worker's work and whether the worker deserved a bonus. Finally, participants provided rationale to justify their decisions and qualitative feedback to help the worker improve.

- *Review the writing task.* Participants made seven Yes/No judgments on whether the product review 1) was an original product review; 2) contained sufficient information about the product; 3) was useful; 4) contained personal stories and anecdotes; 5) listed both good and bad aspects of the product; 6) assessed the product's value given its price; and 7) contained any spelling or grammar mistakes.
- *Review the brainstorming task.* Participants judged whether the worker 1) came up with ten different ideas; 2) described the ideas clearly; 3) came up with high quality ideas; 4) came up with creative ideas; and 5) made any spelling or grammar mistakes.
- *Review the moral judgment task.* Participants judged whether 1) the worker's decision took into account the interest of the basketball team who wanted to retain the player; 2) the worker's decision took into account the interest of the course instructor who wanted to punish the cheating behavior; 3) the worker provided good reasons for the decisions; 4) the worker demonstrated good communication skills; and 5) the worker made any spelling or grammar mistakes.
- *Review the mathematics problem.* Participants judged whether the worker 1) gave the correct answer; 2) provided a clear explanation; and 3) made any spelling or grammar mistakes.
- *Review the summary task.* Participants judged whether the worker 1) provided a summary of no more than three sentences; 2) provided a concise summary; 3) provided an accurate summary; 4) included all the key points; and 5) did not have any spelling or grammar mistakes.

Participants

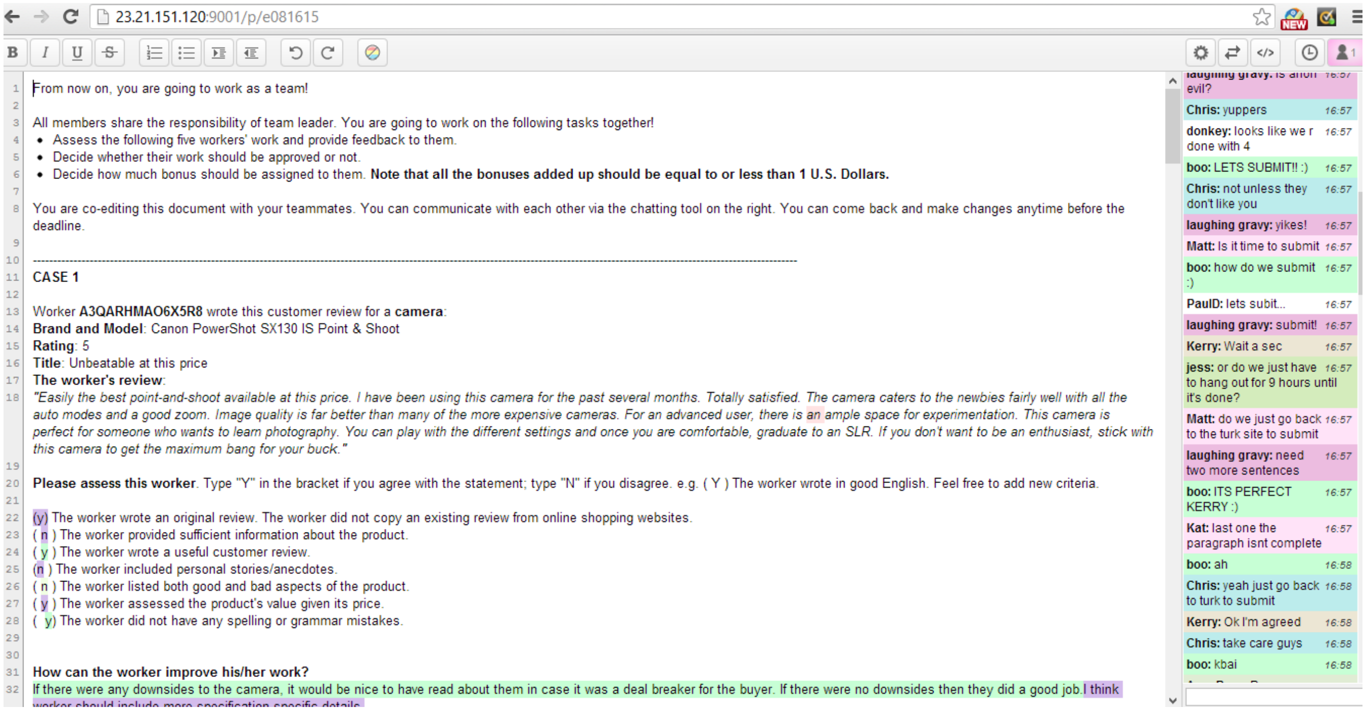
We recruited 680 participants from Amazon Mechanical Turk (MTurk) [12, 24]. To ensure quality, we restricted recruitment to workers with a 90% HIT acceptance rate.

Experiment Procedure

After accepting the MTurk task, participants completed the pre-task. Then we gave participants a link that randomly directed them to an Etherpad that corresponded with one of the ten conditions. When workers finished performing their main tasks on the pad, they went back to MTurk to get paid. Each worker received $1.50 by finishing the pre-task and main experiment. After submitting the main task, workers received a post-task invitation email. Workers received $1 for completing the post-task.

Study Instrument

To perform their work (and to facilitate communication in the interactive team condition) participants used Etherpad-lite, an open-source collaborative editor (see Figure 2). The editor highlights each user's input using a unique color. Users on the same pad can see others' changes in real-time. Users communicate with each other using the chat functionality on the right side. Note that the editor enables participants who arrive asynchronously to collaborate with others: they had access to the entire chat history and the existing output (as a way of interacting with previous work-



23.21.151.120:9001/p/e081615

From now on, you are going to work as a team!

All members share the responsibility of team leader. You are going to work on the following tasks together!

- Assess the following five workers' work and provide feedback to them.
- Decide whether their work should be approved or not.
- Decide how much bonus should be assigned to them. **Note that all the bonuses added up should be equal to or less than 1 U.S. Dollars.**

You are co-editing this document with your teammates. You can communicate with each other via the chatting tool on the right. You can come back and make changes anytime before the deadline.

CASE 1

Worker **A3QARHMAO6X5R8** wrote this customer review for a **camera**:
Brand and Model: Canon PowerShot SX130 IS Point & Shoot
Rating: 5
Title: Unbeatable at this price
The worker's review:
"Easily the best point-and-shoot available at this price. I have been using this camera for the past several months. Totally satisfied. The camera caters to the newbies fairly well with all the auto modes and a good zoom. Image quality is far better than many of the more expensive cameras. For an advanced user, there is an ample space for experimentation. This camera is perfect for someone who wants to learn photography. You can play with the different settings and once you are comfortable, graduate to an SLR. If you don't want to be an enthusiast, stick with this camera to get the maximum bang for your buck."

Please assess this worker. Type "Y" in the bracket if you agree with the statement; type "N" if you disagree. e.g. (Y) The worker wrote in good English. Feel free to add new criteria.

(y) The worker wrote an original review. The worker did not copy an existing review from online shopping websites.
(n) The worker provided sufficient information about the product.
(y) The worker wrote a useful customer review.
(n) The worker included personal stories/anecdotes.
(n) The worker listed both good and bad aspects of the product.
(y) The worker assessed the product's value given its price.
(y) The worker did not have any spelling or grammar mistakes.

How can the worker improve his/her work?
If there were any downsides to the camera, it would be nice to have read about them in case it was a deal breaker for the buyer. If there were no downsides then they did a good job.I think worker should include more specification specific details

laughing gravy: is anon evil? *16:57*
Chris: yuppers *16:57*
donkey: looks like we r done with 4 *16:57*
boo: LETS SUBMIT!! :) *16:57*
Chris: not unless they don't like you *16:57*
laughing gravy: yikes! *16:57*
Matt: Is it time to submit *16:57*
boo: how do we submit :) *16:57*
PaulD: lets subit... *16:57*
laughing gravy: submit! *16:57*
Kerry: Wait a sec *16:57*
jess: or do we just have to hang out for 9 hours until it's done? *16:57*
Matt: do we just go back to the turk site to submit *16:57*
laughing gravy: need two more sentences *16:57*
boo: ITS PERFECT KERRY :) *16:57*
Kat: last one the paragraph isnt complete *16:57*
boo: ah *16:58*
Chris: yeah just go back to turk to submit *16:58*
Kerry: Ok I'm agreed *16:58*
Chris: take care guys *16:58*
boo: kbai *16:58*

Figure 2. Participants used an Etherpad to perform production work and to perform review task (assessing other workers).

ers); and they could leave comments on the pad or on the chat history (to interact with the future workers). We directed participants that were part of the same interactive team to the same pad where they worked together, although sometimes members of a team arrive asynchronously within around half an hour. In contrast, participants in the individual condition worked alone on a single pad.

Measures

To examine the effects of participating in review tasks on reviewers' learning, we compared their performance on the pre-task and post-task. Independent judges blind to condition rated participants' performance on both tasks. For the brainstorming task, judges rated the novelty of the ideas on a scale of 1-5. For the summary task, judges rated the accuracy and conciseness (on a scale of 1-5) of the summary. The change in score from pre-task to post-task indicates the learning gain.

To examine the effectiveness of different review strategies, we calculated five measures about the quality of the assessment task..

- *Consistency among reviewers.* We used intra-class correlation (ICC) to measure how consistently each reviewer or team's assessment agreed with other assessment of the same case. In our experiment ICC is a descriptive statistic that measures how much judgments of the case resemble each other.
- *Internal coherence.* As a measure of the judgment coherence, we calculated the degree to which reviewers' or teams' overall evaluation of a task correlated with their evaluation of more detailed evaluations of the task. For example, were their overall evaluation of a summary predicted by their evaluations of its accuracy and conciseness. We ran a prediction model using the reviewer or reviewer team's Yes/No judgments to predict the same reviewer or reviewer team's decisions on task approval and bonus assignment. Our assumption is that a high coefficient of determination (R-square) indicates that one can predict the final judgment according to its component judgments. The higher the R-square value, the more coherent the assessment.
- *Agreement with judges.* To evaluate the accuracy of participant assessments, we compared them to judgments made by independent assessors — two PhD students not part of our research team. The judges reviewed the tasks together and created an assessment for each. During the main task, reviewers or reviewer teams made a set of assessments, including the Yes/No judgments and the decisions of task approval and bonus assignment. For each reviewer or reviewer team, we transformed their assessment to a vector. For example, one reviewer made the following assessment about a mathematics problem: 1) the worker solved the problem correctly; 2) the worker explained the solution clearly; 3) the work contained some grammar errors; 4) the worker's work should be approved; and 5) the worker does not deserve a bonus. The vector for this reviewer's assessment is (1,1,0,1,0). Then we calculated the distance between the assessments of participants and the independent judges. The shorter

the distance, the more the reviewer (team) agreed with the judges. We calculated the "agreement with the judge" as the reverse distance.

- *Subjective ratings of helpfulness.* Reviewers provided a rationale for their decision about task approval and bonus assignment, and qualitative feedback to help workers to improve their work. To judge the quality of the assessment rationale and qualitative feedback, we recruited a new set of independent Mechanical Turk workers blind to condition to rate the convincingness of the rationale and the usefulness of the feedback on five-point Likert scales.
- *Correctness of judging mathematics problem.* We measured whether the reviewers or reviewer teams made the right judgment about the correctness of the mathematics problem.

Data Preparation

Of the 680 participants recruited for the experiment, half were placed in reviewer conditions and the other half in producer conditions. There were 40 individuals and 40 interactive teams each condition (see Table 1).

Table 1 shows the participation numbers per condition for the pre/main task and post-task. We found no significant difference on the number of people returning for the post-task between reviewer conditions and producer conditions ($\chi^2(1) = 1.40$, p=0.24).

To assess the quality of reviews for different conditions, we include all the reviews from teams and individuals in the analysis and treated incomplete items as missing data. The individual completion rate is higher for reviewers in individual conditions. The dropout rate (i.e., percentage of people did not submit the task) of individual reviewer is 17.5%, while the dropout rate of participants in interactive team review condition is 38% ($\chi^2(1) = 6.47$, p=0.01). However, in terms of the completion of review task as a team (i.e., completing all the assessment items), individual reviewer is 83.3% while review teams are 92%.($\chi^2(1) = 51.9$, p<0.01).

		# of Teams	# of participants (pre/ main task)	# of participants (post-task)	Return rate
Reviewer	Individual	NA	40	16	40%
	Team size 2	10	20	9	45%
	Team size 4	10	40	13	33%
	Team size 8	10	80	30	38%
	Team size 16	10	160	54	34%
Producer	Individual	NA	40	11	28%
	Team size 2	10	20	6	30%
	Team size 4	10	40	12	30%
	Team size 8	10	80	38	48%
	Team size 16	10	160	70	44%

Table 1. Participation numbers per condition for the main task and post-task

Construction of Nominal Teams

To understand if teams should be organized by simply pooling individual efforts, we formed nominal teams by combining the decisions of the individual workers. Specifically, to construct a nominal team of N workers, we randomly selected N individual workers who reviewed the same case and combined their individual outputs as the nominal team output. For Yes/No judgments, we selected the majority opinion as the team judgment. If there was a tie, we randomly selected Yes or No as the output. For qualitative outputs such as rationale for the decisions and feedback, we concatenated the individual output to form the team output. We constructed 10 nominal teams for each team size (2, 4, 8, and 16). Note that nominal teams reuse the data from individual conditions. Therefore, we will not directly compare these two conditions in the analysis since they are interdependent. Our analyses only draw comparison between interactive teams and *either* individuals or nominal teams.

RESULTS

Effect of Reviewing on Task Performance

We found that workers who participated in reviewing had higher learning gain (comparing pre- vs. post-task results) than those who did production work, which supports our hypothesis 1.

Reviewers Learned More

We measured learning gains by comparing participants' task performance on a pre- and post-tasks. For independent variables, we created three dummy variables indicating whether it is pre-task or post task, whether the participant was in the reviewer conditions or in the producers' conditions, and whether the participant worked individually or in groups. We only included participants who finished both pre-task and post-task in the analysis.

	DV: novelty of the ideas in the brainstorming task		
	Model 1	Model 2	Model 3
	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Pre-Post (0- pre; 1 -post)	0.13 (0.04)**	-0.07 (0.06)	0.01 (0.18)
Producer–Reviewer (0-producer; 1-reviewer)	-0.02 (0.04)	0.09 (0.05)	0.10 (0.20)
Pre-Post X Producer-Reviewer	NA	0.11 (0.05)*	0.27 (0.17)
Individual-Team (0-individual, 1-team)	NA	NA	0.07 (0.15)
Pre-post X Individual-Team	NA	NA	0.16 (0.14)
Producer-Reviewer XIndividual-Team	NA	NA	-0.03 (0.15)
Pre-post X Producer-Reviewer X Individual-Team	NA	NA	0.15 (0.15)

Table 2. Linear regression models examining the learning gains. All participants improved from pre to post task (p<.01); Reviewers improved more than producers (*p<.05).**

Model 1 in Table 2 shows that participants' brainstorming performance improved significantly from the pre- to post-test (b=.13, *p*<.05). Model 2, which includes the interaction between "Pre-Post" and "Producer-Reviewer", shows that only participants in the reviewer conditions improved. Performance of those in the producer conditions declined slightly (b= -.07, *p*> .10), while those in reviewer condition significantly improved (b=.11, *p*<.05). There was no significant difference between group conditions and individual conditions in brainstorming task (Model 3, Table 2).

There is a significant improvement in performance ratings between pre- and post- versions of the summary task (for accuracy: b=.27, se=.07, *p*<.01; for conciseness: b=.13, se=.06, *p*<.05), but the increase is not significantly greater for reviewers than producers (for the interaction of accuracy, b=.15, se=14, *p*=.31; for the interaction of conciseness, b=.02, se=13, *p*=.88). We also found no effect of team size on summary task performance.

Effect of Organizational Structures on Review Quality

We found interactive reviewer teams outperformed individual reviewers on all quality measures. Nominal groups outperformed interactive teams on all measures except the correctness of the mathematics problem. However, the gap between nominal teams and interactive teams does not increase as the team becomes larger.

Teams Showed More Consistency

The intra-class correlation among individual reviewers is 0.33; the intra-class correlation among interactive reviewer teams is 0.45; and the intra-class correlation among nominal reviewer teams is 0.61. We used Fisher r-to-z transformation and found that the difference between individual reviewers and interactive reviewer teams is marginally significant (z=1.4, *p* =0.08) and the difference between interactive teams and nominal teams is significant (z = 2.23, p<0.05).

Nominal Team Assessments Were More Coherent

As described previously, to calculate internal coherence we ran a prediction model using the Yes/No judgments to predict the decisions on task approval and bonus assignment on different assessment tasks. Therefore, the higher the R-square value, the more likely the assessment is coherent. The average R-square of nominal teams is 0.66; and interactive teams is 0.6, and the individual condition is 0.56. (see Table 3). We applied Fisher r-to-z transformation on R (the root square of R square) and found that the difference between individual team and interactive team is not significant (z =0.69, p = 0.25); and the difference between interactive team and nominal team is marginal significant (z = 1.35, p=0.09).

Team Assessments Aligned with Judge Assessments

To evaluate the accuracy of participant assessments, we compared them to those of independent judges. Model 1 in Table 4 shows the interactive reviewer team's assessment agrees with the independent judges more than the individual reviewers (b=.05, *p*<.05); and the nominal reviewer team's

	R-square		
Tasks	**Individual reviewers**	**Interactive reviewer teams**	**Nominal reviewer teams**
Writing	0.36	0.45	0.67
Brainstorming	0.76	0.78	0.73
Moral judgment	0.60	0.80	0.90
Mathematics	0.57	0.80	0.43
Summary	0.49	0.17	0.57
Average	0.56	0.6	0.66

Table 3. Internal coherence of the assessment. Nominal teams were more internally coherent.

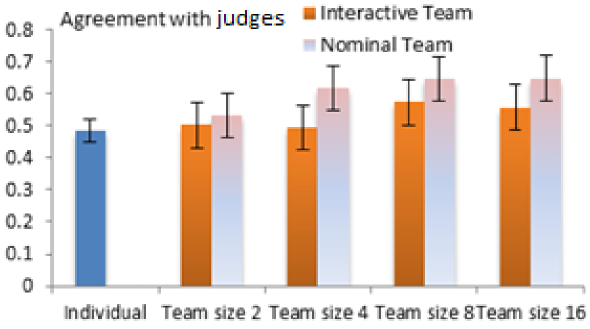


Figure 5. Agreement with judges

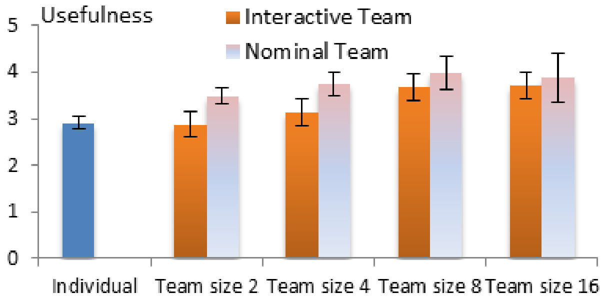


Figure 6. Turkers' rating: perceived usefulness of the feedback

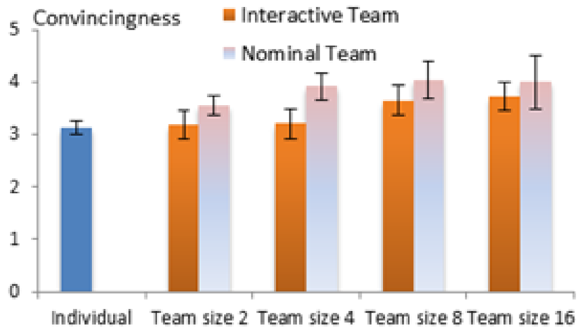


Figure 7. Turkers' rating: perceived convincingness of the rationale.

judgment is even closer than interactive teams (b=.08, *p*<.01). From Model 1 of Table 5, we found that the assessment of larger groups is not significantly closer to the independent judges (b= .025, *p*>.10) compared to smaller groups. There is no significant interaction between team size the team type (interactive team versus nominal team). We can see the trend graphically in Figure 5.

Teams Provided More Useful Feedback

Results show that nominal teams provide higher quality rationale and more useful feedback than interactive reviewer teams (see Figure 6 and 7, and Table 4 for p-values). Moreover, interactive reviewer teams significantly outperform individuals. Compared to the individual reviewers, the interactive reviewer teams score 0.45 points higher for perceived usefulness (b=.45; *p*<.01 in the Model 2 of Table 4) and 0.31 points higher for perceived convincingness (b=.31; *p*<.05 in the Model 3 of Table 4) on a five-point Likert scale; nominal teams score another 0.29 points higher for perceived usefulness than interactive teams (b=.29; *p*<.01 in the Model 2 of Table 4) and 0.30 points higher for perceived convincingness (b=.30; *p*<.01 in the Model 3 of Table 4). Furthermore, ratings of larger groups are higher than ratings of smaller groups. For perceived usefulness, the coefficient of the logarithm of team size is 0.30, *p*<.01; for perceived convincingness, the coefficient is 0.21, *p*<.01 See Model 2 and Model 3 in Table 5).

However, the interaction between team size and team type is not significant, indicating that the gap between nominal teams and interactive teams did not increase as team size grows.

Interactive Teams Exert Coordination Effort

Consistent with our hypotheses, interactive reviewer teams exhibited "production loss". Compared to nominal groups, interactive groups had to devote time to interaction and coordination with other team members. For example, in one Etherpad (see Figure 8), "4kwood" expressed different opinions about how to coordinate the work with the rest of the participants. 4kwood wanted the team to work on the task simultaneously, while the other team members wanted to divide the task first and then review the task together at the end. Feeling ignored, 4kwood quit the task and later sent the following message to the research team:

> *"I just returned a task I was working on. A few of the other workers were somewhat offensive and were more concerned about how much they did rather than working as a team. The task became a race to some participants with quite a bit of sarcasm added in as well."*

Interactive Teams Judged Mathematics Problem Best

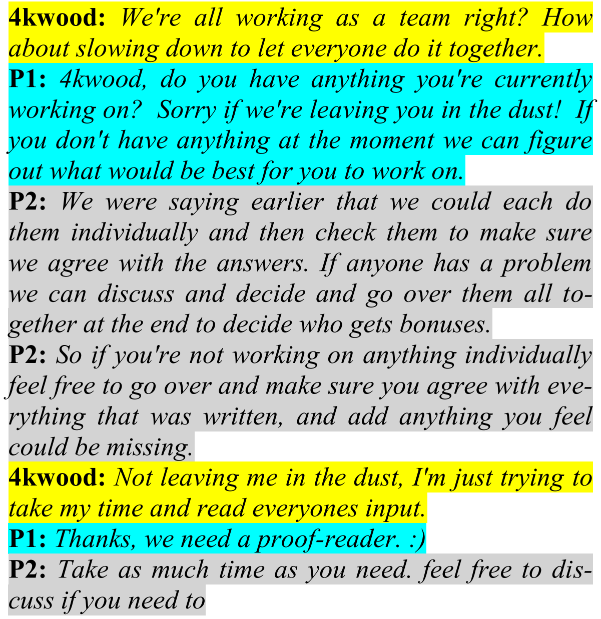
One exception to the trend thus far was that interactive reviewer teams outperformed nominal teams on judging the correctness of the mathematics problem. Specifically, 31% of individual reviewers correctly judged the mathematics problem; 58% of the interactive reviewer teams judged the math problem correctly; and 47% of the nominal teams judged the math problem correctly. The relationship be-

	Model 1	**Model 2**	**Model 3**
	DV: Agreement with judges	**DV: Perceived usefulness**	**DV: Perceived convincingness**
	Coef. (S.E.)	**Coef. (S.E.)**	**Coef. (S.E.)**
Interactive-Individual (1-interactive; 0- individual)	0.05* (0.03)	0.45 ** (0.10)	0.31** (0.10)
Nominal-Interactive (1-nominal; 0-interactive)	0.08** (0.03)	0.29 ** (0.10)	0.30** (0.10)

Table 4. Linear regression models examining the agreement with judges, perceived usefulness and perceived convincingness for individual reviewers, interactive reviewer teams and nominal reviewer teams.*p<.05; **p<.01

	Model 1	**Model 2**	**Model 3**
	DV: Agreement with judges	**DV: Perceived usefulness**	**DV: Perceived convincingness**
	Coef. (S.E.)	**Coef. (S.E.)**	**Coef. (S.E.)**
Nominal-Interactive (1-nominal; 0-interactive)	0.04 (0.08)	0.85** (0.28)	0.47 (0.28)
Logarithm of team size to base two	0.025 (0.015)	0.30** (0.06)	0.21** (0.04)
Nominal-Interactive X team size	00.01 (0.02)	-0.12 (0.09)	-0.001 (0.09)

Table 5. Linear regression models examining the agreement with judges, perceived usefulness and perceived convincingness for different types of teams and different team size.*p<.05; **p<.01



4kwood: *We're all working as a team right? How about slowing down to let everyone do it together.*

P1: *4kwood, do you have anything you're currently working on? Sorry if we're leaving you in the dust! If you don't have anything at the moment we can figure out what would be best for you to work on.*

P2: *We were saying earlier that we could each do them individually and then check them to make sure we agree with the answers. If anyone has a problem we can discuss and decide and go over them all together at the end to decide who gets bonuses.*

P2: *So if you're not working on anything individually feel free to go over and make sure you agree with everything that was written, and add anything you feel could be missing.*

4kwood: *Not leaving me in the dust, I'm just trying to take my time and read everyones input.*

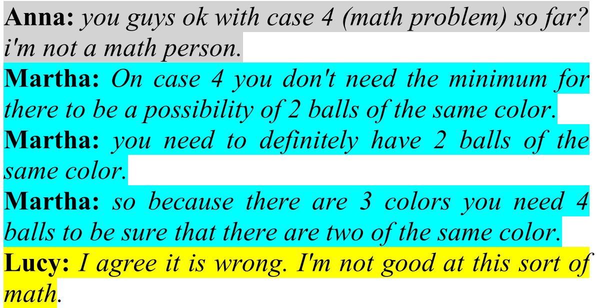
P1: *Thanks, we need a proof-reader. :)*

P2: *Take as much time as you need. feel free to discuss if you need to*

Figure 8. Conversations demonstrating the production

tween the correctness and type of review structure is significant, chi-square with two degrees of freedom is 6.26, $p<.05$.

One reason for this could be that the probability problem was a “truth supported” task [22]. In such tasks, group members may differ in their ability to solve a problem, but can recognize a good solution once a group member had solved it. Since fewer than half of the individual reviewers judged the mathematics problem correctly, the nominal group’s structured algorithm to pool the individual results does not effectively select the correct judgment. In other words, individuals in the nominal groups never see their group members' answers, so they never have the opportunity to get struck by the insight. One benefit of interactive teams is that individuals who know the correct answer can influence the whole team through discussions [17]. Specifically, participants who understood the mathematics problem could convince teammates and influence the final judgment, even when they are in the minority. For example, one participant (Martha) successfully convinced other members in the team and led the team to reach the correct judgment on the mathematics problem (see Figure 9).



Anna: *you guys ok with case 4 (math problem) so far? i'm not a math person.*

Martha: *On case 4 you don't need the minimum for there to be a possibility of 2 balls of the same color.*

Martha: *you need to definitely have 2 balls of the same color.*

Martha: *so because there are 3 colors you need 4 balls to be sure that there are two of the same color.*

Lucy: *I agree it is wrong. I'm not good at this sort of math*.

Figure 9. Conversations on mathematics problem.

DISCUSSION

In this paper, we explored the beneficial effects of reviewing others’ work on learning and performance. We found that workers who participated in review tasks had higher learning gains than workers who only did production tasks.

We also examined how to organize people into teams and found that interactive reviewer teams outperformed individual reviewers on most quality measures including consistency among reviewers, agreement with judges, subjective ratings of the assessments’ helpfulness to workers, and correctness of judging mathematics problems. Moreover, nominal reviewer teams performed even better than interactive teams on all the measures except the correctness of mathematics problem. Qualitative evidence indicates that interactive teams exerted extra coordination efforts. Also, some discussions resulted in unhealthy social influence causing people to move away from the “truth” and towards each other’s initial biased positions. With exception of the math problem where discussion may have helped to overcome individual misconceptions, the observations of coordination costs and social biasing may explain why the assessments produced by interactive teams were less consistent, were less coherent, and had lower agreement with independent judges.

Our results showed that the performance gaps between nominal teams and interactive teams were not larger as team size increased. It is possible that the factors inhibiting the effectiveness of interactive teams might have some ceiling effects.

Design Implications

The results demonstrate that providing opportunities to review others can improve workers’ learning. In addition, outsourcing the review task should also have a downstream benefit for the production workers who receive the assessment, as shown in previous work [5]. Although we did not explicitly measure the downstream benefit, a separate set of Turkers judged whether the reviewers provided helpful feedback in their assessments. The average score of larger nominal teams (>2) is around 4 in a 5 point scale. Therefore, we have reasons to believe that assessment and feedback generated by peer reviewers should help the production workers improve skills and task performance. In theory, we claim that outsourcing review tasks can facilitate the learning of both reviewers (who assess others’ work) and production workers (who receive assessment), and thus benefit the employers to accomplish higher quality work in the long run.

The study also provides guidelines for how employers might structure the review task. For subjective tasks, we suggest employers first allow reviewers to do the assessment task individually and then systematically aggregate their assessment (a nominal team approach). For more objective tasks in which correct answers can “win” through rational discussion (like the Mathematics problem in our study), employers should consider providing communication channels for reviewers so that they can identify the correct judgment through discussion.

Limitations and Future Work

We measured learning gain by comparing workers’ pre- to post performance on only two tasks (i.e., brainstorming and summary), and in a specific way (comparing the novelty of the ideas and the accuracy and conciseness of the summaries from pre-task to post-task). It is possible that by reviewing others, workers may have been learning in ways that did not show up on our tests. For example, workers might have improved their task performance on other types of tasks, they might have learned review techniques such as how to provide useful feedback, or they might have learned communication and coordination strategies by working with others. One direction of future work is to further explore what else workers could learn from reviewing others.

Factors such as coordination and free-riding cause production loss and undermine the performance of interactive teams. However, working in interactive teams might have other benefits. For example, Sutton and Hargadon studied a product design firm and found that interactive brainstorming sessions can support organizational memory of design solutions, help members recognize skill variety among team, and build shared ownership of ideas [31]. In our experiment, we also observed that workers appreciate the experience of working together with others. In future work, we plan to measure whether working in interactive teams improves non-quality outcomes such as work satisfaction and loyalty.

We constructed nominal teams using a simple linear algorithm (majority algorithm). Small group research shows that the majority algorithm is better than a single judge (where a single person aggregates all the team members' opinions and makes the final judgment) [1] and other more stringent algorithms (e.g., two-thirds majorities or unanimity) [11]. Our results showed that nominal teams using the majority algorithm did perform best on most quality measures except on judging mathematics problems. It is possible that we can improve nominal teams' performance on these types of tasks by improving the combining algorithm. For example, we can ask each member to assign a confidence parameter for their own judgment and set this parameter as weight when combining members' judgments.

CONCLUSION

Our paper demonstrates that providing workers opportunities to review other workers' work will enhance their understanding of the task domain, and eventually help them become better workers. We investigated the effectiveness of three different strategies for organizing workers to accomplish review tasks: assigning individual reviewers, organizing interactive reviewer teams, and aggregating individual reviewers into nominal teams. Our results provide practical guidelines for employers and crowdsourcing system designers to better structure and train the workforce to accomplish high-quality work.

ACKNOWLEDGMENTS

This work was supported by NSF grants IIS-1149797, IIS-1217559, OCI-0943148, IIS-0968484, IIS-1111124, IIS-1208382, IIS-1217096, Bosch, Google, and Microsoft.

REFERENCES

1. Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, et al. 2000. The effects of averaging subjective probability estimates between and within judges. J. Exp. Psychol.: Applied 6:130–47.
2. Barry, D.(1991) Managing the bossless team: Lessons in distributed leadership. Organizational Dynamics, 20, 31-47.
3. Brown, A., Bransford, J., Ferrara, R.,& Campione, J.(1983). Learning, remembering, and understanding. In J.H. Flavell & E.M. Markman (Eds.), Handbook of child psychology: Vol.3. Cognitive development (4th ed., pp.77-166). New York: Wiley.
4. Chase C.C., Chin, D.B., Oppezzo, M.A., and Schwartz, D.L. (2009) Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. Journal of Science Education and Technology. Volume 18, Number 4 (2009), 334-352.
5. Dow, S.P., Kulkarni, A., Klemmer, S.R., and Hartmann, B.(2012) Shepherding the Crowd Yields Better Work. ACM Conference on Computer Supported Cooperative Work, 2012.
6. Felps, W., Mitchell, T., and Byington, E. How, When, and Why Bad Apples Spoil the Barrel: Negative Group Members and Dysfunctional Groups. Research in Organizational Behavior 27, (2006), 175-222.
7. Guilford, J. P. (1967). The nature of human intelligence. New York, NY: McGraw Hill.
8. Horton, J. J. (2010). Employer expectations, peer effects and productivity: Evidence from a series of field experiments. arXiv preprint arXiv:1008.2437.
9. Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. Journal of Personality and Social Psychology, 50(6), 1141.
10. Janis, I.L. Groupthink: Psychological Studies of Policy Decisions and Fiascoes. Wadsworth Publishing, 1982.
11. Kerr, N. L., Tindale, R.S. (2004) Group performance and decision making. Annu. Rev. Psychol. 2004. 55:623–55.
12. Kittur, A., Chi, E.H., and Suh, B. (2008) Crowdsourcing user studies with Mechanical Turk. In Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08). ACM, New York, NY, USA, 453-456.
13. Kittur, A. and Kraut, R.E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08)
14. Kittur, A., Smus, B., Khamkar, S., and Kraut, R.E. (2011) CrowdForge: crowdsourcing complex work. In Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11). ACM, New York, NY, USA, 43-52.
15. Kraut, R. E. & Resnick, P. (2012). Building successful online communities: Evidence-based social design. Cambridge, MA: MIT Press.
16. Kulkarni, A., Gutheim, P., Narula, P., Rolnitzky, D., Parikh, T., & Hartmann, B. (2012). MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture. IEEE Internet Computing, 16(5), 28-35.
17. Laughlin PR, McGlynn RP. (1986). Collective induction: mutual group and individual influence by exchange of hypotheses and evidence. J. Exp. Soc. Psychol. 22:567–589.
18. Levine, J.M., Moreland, R.L. (1990) Progress in small group research. Annual Review of Psychology, Vol 41, 1990, 585-634.
19. Mannix, E., and Neale, M.A. (2005) What Differences Make a Difference? The Promise and Reality of Diverse Teams in Organizations. Psychological Science in the Public Interest , Vol. 6, No. 2 (Oct., 2005), pp. 31-55.
20. Matsuda, N., Cohen, W. W., Koedinger, K. R., Keiser, V., Raizada, R., Yarzebinski, E., et al. (2012). Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations. In Digital Game and Intelligent Toy Enhanced Learning (DIGITEL), 2012 IEEE Fourth International Conference on (pp. 25-32). IEEE.
21. Manz, C.C., & Sims, H.P., Jr(1993) Business without bosses. New York: Wiley.

22. McGrath, J.E. (1984) Groups: interaction and performance. Prentice Hall, Inc., Englewood Cliffs. N.J.

23. Mullen, B., Johnson C., and Salas E.(1991) "Productivity loss in brainstorming groups: A meta-analytic integration." Basic and Applied Psychology, 12: 2-23.

24. Paolacci, G., Chandler, J., Ipeirotis, P.G. (2010) Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, Vol 5(5), Aug 2010, 411-419.

25. Paulus, P.B., Dzindolet, M.T., Poletes, G., and Camacho, L.M. (1993) "Perceptions of performance in group brainstorming: The illusion of group productivity." Personality and Social Psychology Bulletin, 19: 78-89.

26. Pearce, C.L., and Conger,J. A. (2003) Shared leadership: Reframing the hows and whys of leadership, Sage, Thousand Oaks.

27. Ryan, R. M.; Deci, E L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, Vol 55(1), Jan 2000, 68-78.

28. Schkade, D., Sunstein, C. R., & Kahneman, D. (2000). Deliberating about dollars: The severity shift. Columbia Law Review, 1139-1175.

29. Srivastava A., Bartol, K.M., and Locke, E.A.(2006) Empowering Leadership in Management Teams: Effects on Knowledge Sharing, Efficacy, and Performance. The Academy of Management Journal , Vol. 49, No. 6 (Dec., 2006), pp. 1239-1251.

30. Stroebe, W. and Diehl, M. Why Groups are less Effective than their Members: On Productivity Losses in Idea-generating Groups. European Review of Social Psychology 5, (1994), 271.

31. Sutton, R. and Hargadon, A. Brainstorming groups in context: effectiveness in a product design firm. Administrative Science Quarterly, (1996).

32. Van De Ven, A. H., and Delbecq A.L. (1971) Nominal versus Interacting Group Processes for Committee Decision-Making Effectiveness. The Academy of Management Journal, Vol. 14, No. 2 (Jun., 1971), pp. 203-212.

33. Van De Ven, A. H., and Delbecq A.L. (1974) The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes. The Academy of Management Journal, Vol. 17, No. 4 (Dec., 1974), pp. 605-621.

34. White, B.Y. and Frederiksen, J.R. (1998) Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. Cognition and Instruction, Vol. 16, No. 1 (1998), pp. 3-118.

35. Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., Malone, T.W. (2010) Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science 29 October 2010: Vol. 330 no. 6004 pp. 686-688

36. Yukl, G. A. (1998). Leadership in organizations (4th ed.). Upper Saddle River, N.J.: Prentice Hall.

37. Zhu, H., Kraut, R.E., & Kittur, A., (2012) Effectiveness of Shared Leadership in Online Communities. CSCW'2012.

38. Zhu, H., Kraut, R., & Kittur, A. (2012, February). Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 935-944). ACM.

39. Zhu, H., Kraut, R.E., Wang, Y.C., & Kittur, A. (2011) Identifying Shared Leadership in Wikipedia. In CHI'2011: Proceedings of the 2011 annual conference on Human factors in computing systems. New York: ACM Press.

.