Paying Crowd Workers for Collaborative Work

GREG D'EON, University of Waterloo, Canada
JOSLIN GOH, University of Waterloo, Canada
KATE LARSON, University of Waterloo, Canada
EDITH LAW, University of Waterloo, Canada

Collaborative crowdsourcing tasks allow crowd workers to solve problems that they could not handle alone, but worker motivation in these tasks is not well understood. In this paper, we study how to motivate groups of workers by paying them equitably. To this end, we characterize existing collaborative tasks based on the types of information available to crowd workers. Then, we apply concepts from equity theory to show how fair payments relate to worker motivation, and we propose two theoretically grounded classes of fair payments. Finally, we run two experiments using an audio transcription task on Amazon Mechanical Turk to understand how workers perceive these payments. Our results show that workers recognize fair and unfair payment divisions, but are biased toward payments that reward them more. Additionally, our data suggests that fair payments could lead to a small increase in worker effort. These results inform the design of future collaborative crowdsourcing tasks.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: crowdsourcing, incentives, equity theory, cooperative game theory

ACM Reference Format:
Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying Crowd Workers for Collaborative Work. *Proc. ACM Hum.-Comput. Interact.* 3, 1, Article 1 (November 2019), 24 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Micro-task crowdsourcing platforms, such as Amazon Mechanical Turk, allow requesters to hire human workers to complete short, self-contained tasks. These tasks are typically meant to be completed individually: workers might label images or transcribe audio clips on their own. However, collaborative tasks can solve new problems by relying on contributions from multiple workers. One approach is to break a difficult problem into a workflow of simpler steps [3, 29]. Another is to have workers justify and debate their answers in a structured manner [5, 10, 58]. Some tasks even allow free-form communication between workers, allowing them to brainstorm or cooperate on complex intellectual problems [40, 56, 65]. These techniques allow crowdsourcing systems to solve difficult problems by enabling interactions between multiple workers.

Collaborative crowdsourcing tasks, however, introduce new challenges in motivating workers. Workers are primarily motivated by money [25], and the implications of this motivation have been thoroughly studied for individual tasks. Higher pay attracts workers more quickly [54] and

Authors' addresses: Greg d'Eon, greg.deon@uwaterloo.ca, University of Waterloo, Waterloo, Canada; Joslin Goh, jtcgoh@uwaterloo.ca, University of Waterloo, Waterloo, Canada; Kate Larson, klarson@uwaterloo.ca, University of Waterloo, Waterloo, Canada; Edith Law, edith.law@uwaterloo.ca, University of Waterloo, Waterloo, Canada.

2573-0142/2019/11-ART1 $15.00
https://doi.org/0000001.0000001

causes them to complete more work [46], while performance-based bonuses can increase worker effort [18]. However, in collaborative tasks, workers can often see each others' work, and this extra information may have a large impact on their motivation. For instance, the simplest payment strategy—to pay all workers equally—may not be suitable, as the most skilled workers could feel undervalued if they know others are earning the same wages.

More precisely, in collaborative work, workers can compare against others in their group to judge whether their payments are equitable. Equity theory [1] posits that people believe their rewards should be proportional to the quality or quantity of their work or the time they spend on the job. When people are paid too much or too little, they often restore the equity balance by putting more or less effort into their work. These predictions have been verified in laboratory studies [51] and using real-world salary data [16]; however, our work is the first to validate equity theory on crowdsourcing platforms.

In this paper, we examine the problem of paying groups of crowd workers for collaborative work. First, we review existing collaborative crowdsourcing tasks and categorize these tasks into distinct styles of collaboration. Then, we propose two theoretically fair payment methods—proportional payments based on equity theory and the Shapley value from cooperative game theory [4]—and discuss how these payments can motivate small groups of crowd workers. We conduct two user studies to evaluate the practical impacts of these payment methods. In the first study, we hire crowd workers for a collaborative task and compare their perceptions of fairness when they are paid equal, proportional, or Shapley-valued bonuses. In the second, we ask a separate set of impartial crowd workers to evaluate the fairness of these payments. We show that workers perceive these theoretically grounded payments as being more fair, but are biased toward payments that reward them more. Our results also suggest that workers exert similar amounts of effort regardless of the payment method. Finally, we use our results from these studies to make recommendations about rewards in future collaborative crowdsourcing systems.

This work makes three key contributions to the crowdsourcing literature:

- We provide a categorization of existing collaborative crowdsourcing tasks.
- We describe the connection between worker motivation and fair payments in collaborative tasks, and we propose two payment methods that are grounded in equity theory and cooperative game theory.
- We present empirical results showing the impacts of these theoretically fair payments in a realistic crowdsourcing task.

In the remainder of this paper, we review the existing literature on collaborative crowdsourcing tasks, discuss the theoretical link between motivation and fair payment divisions, describe our experiments and results, and conclude with specific recommendations for future requesters designing collaborative tasks.

2 COLLABORATIVE CROWDSOURCING TASKS

We begin by reviewing the existing literature on collaborative crowdsourcing tasks. We use this literature review to identify the different types of information that are available to workers during collaborative tasks. These features help to identify a number of distinct categories of collaborative work, each embodying a different level of interaction between the workers.

To define what we mean by collaborative tasks, we follow Malone and Crowston [42], who define collaboration as "peers working together on an intellectual endeavor". Based on this, we take collaborative crowdsourcing to include *any crowdsourcing task where work from multiple workers is used to produce a single result.* Note that this is quite a broad definition: for example, it includes

systems where answers from independent workers are aggregated without any interaction between the workers.

Note that collaborative tasks can also be competitive. To be precise, Malone and Crowston [42] state that cooperation indicates situations where actors share the same goals, while competition connotes one actor gaining from another's losses. Group work typically includes both of these elements: Davis [9] notes the extremes of pure cooperation or pure competition are rare. Most group-based crowdsourcing tasks also fall into quadrants 1 ("generate") and 2 ("choose") of McGrath's task circumplex [47]. While tasks in these quadrants are primarily cooperative, they also include elements of competition.

2.1 Dataset and Method

We performed our literature review using a snowball sampling process, a standard procedure for literature reviews [36]. Our search was seeded with Bernstein et al.'s Soylent [3]: as one of the first crowdsourced workflows, it represents one of the earliest and most recognized collaborative tasks. Then, we iteratively reviewed references in both directions by checking the reference lists and Google Scholar "cited by" lists. We kept all papers that described a collaborative crowdsourcing task. This process resulted in a total of 114 papers. The majority of these papers describe tasks for Mechanical Turk, with a small number focusing on professional crowdsourcing (e.g., Upwork) or citizen science (e.g., Zooniverse) platforms.

We used an iterative coding process to analyze these collaborative tasks. Our first round of coding began with a subset of 40 papers. We analyzed these papers by looking for features in the task descriptions and interfaces that showed how workers collaborated during their work. In this first round, we converged on 4 features that differentiate these collaborative tasks from each other. Each of these features describes one type of information that workers might have available to them during their tasks:

- *See others' work:* Does the task interface include any information showing work done by other workers? If so, was this work done on the *same* task, or on a *different* task?
- *Aware of others:* Does the task description or interface indicate that other workers are involved in the task?
- *Identify others' work:* If they can see others' work, is this information shown with identifiers such as usernames or pseudonyms, or is it anonymous?
- *Freely interact:* Does the task interface allow them to have open, free-form conversations with other workers?

We minimized the amount of ambiguity in these features by phrasing them as answers to a series of binary questions. We then applied this categorization to all of the papers in our sample, iterating on these feature definitions to resolve ambiguous cases when necessary.

2.2 Results

The categories that we discovered are shown in Table 1. We identified four types of collaboration that are relatively common, appearing in at least 10 publications. Characteristics and representative tasks for each of these categories are:

- *No information about others:* Tasks that require input from multiple workers, but do not have any form of interaction between the workers. This category includes most answer aggregation systems. It also includes real-time crowdsourcing tasks such as Adrenaline [2], where workers complete tasks simultaneously with no information about each other.
- *Workflows with no awareness:* Each worker's job depends on data from previous workers, but the data's source is not mentioned. For example, in the final step of Soylent's find-fix-verify

	See others' work	*Aware of others*	*Identify others' work*	*Freely interact*	# papers	Category
Common (10+ papers)	N	N	N	N	30	No information about others
	D	N	N	N	15	Workflows (no awareness)
	S	Y	N	N	18	Shared interfaces (anonymous)
	S	Y	Y	Y	17	Full collaboration
Uncommon (0-9 papers)	S	N	N	N	6	Iterative tasks
	N	Y	N	N	3	Aware of other workers
	D	Y	N	N	7	Workflows (with awareness)
	D	Y	Y	N	1	Subcontracting
	S	Y	Y	N	5	Structured deliberation; shared interfaces
	S	Y	N	Y	3	Anonymous chat
Not MTurk	N	Y	N	Y	1	Solo work with chat room
	D	Y	N	Y	3	Workflows with chat
	D	Y	Y	Y	4	Professional workflows

Table 1. The categories of collaborative crowdsourcing tasks that we found in our literature review. For the *See others' work* feature, workers can see others' work for the same task (S), a different task (D), or not at all (N). For the other three features, the collaboration is either present (Y) or not (N).

workflow [3], workers are asked to confirm writing quality without being told that the sentences were rewritten by other Turkers.

- *Anonymous shared interfaces:* Workers contribute to a common, shared interface, but cannot directly communicate or identify which workers performed each part of the work. This approach has been used to control arbitrary GUIs [35], plan complex itineraries [64], and write creative stories [27].
- *Full collaboration:* Workers closely interact as a group. Typically, this type of collaboration is achieved using a shared writing space, such as Google Documents or Etherpads, or using a chatroom such as a Slack workspace. This type of task is often associated with creative thinking [40], complex problem solving [65], or deliberation [6, 58].

We also identified six types of collaboration that are less common in previous work:

- *Iterative tasks:* A series of workers perform the same task, but are given previous results as a starting point or for inspiration. This approach works well for image segmentation [24, 26] and some types of brainstorming [37, 59].
- *Aware of other workers:* The task interface mentions that other workers are completing the same task, but does not show their work. This technique is used to motivate workers in tasks that otherwise consist of individual work [19, 60].
- *Workflows with awareness of workers:* This category includes workflows where the presence of previous workers is explicitly mentioned [14, 28]. It also includes divide-and-conquer workflows [30, 31], where workers decide how complex tasks should be divided.

- *Subcontracting:* Morris et al. [50] proposed a workflow where workers choose to divide complex tasks through "subcontracting", using real-time chat to facilitate assistance between workers.
- *Structured deliberation and shared interfaces:* Some deliberation workflows only allow specific, structured communication between workers [5, 38]. Additionally, in some shared interfaces, it is possible for workers to see what each member of group is doing [21, 34].
- *Anonymous chat:* A small number of tasks involving chat interfaces show all messages coming from the anonymous "crowd" user [20].

Finally, we noted three other styles of collaboration that appear on other platforms, but have not appeared in microtask crowdsourcing. These three categories allow workers to communicate with each other, but vary the amount of cooperative work that they are involved in.

2.3 The Value of Collaboration

Not all tasks require collaboration. For example, structured workflows can be inefficient for simple tasks, as they can increase redundancy and hide important context. However, collaborative work can be valuable, as it allows non-expert workers to tackle complex problems.

Collaboration allows workers to form more effective groups. Olson and Olson [53] described several affordances that allow colocated teams to perform tightly coupled work. Three of these affordances—coreference, personal information, and rapid feedback—are closely aligned with our *see others' work*, *identify others' work*, and *freely communicate* features. Further, with closer interactions, workers can carry out Malone and Crowston's coordination processes [41], rather than relying on requesters to manage the work. For example, in most workflows, requesters must *a priori* break a task into microtasks, while in Apparition [34], workers can decide how to divide their work on the fly. These theoretical connections suggest that highly collaborative crowdsourcing tasks allow for workers to carry out tightly coupled, dynamic work.

For concrete evidence of these advantages, we point out four types of problems that have been solved using structured deliberation, shared interfaces, or full collaboration. First, while individual workers are capable of some simple creative tasks, several creative writing tasks depend on workers having open discussions with each other [40, 56]. Second, workers are better at solving difficult cognitive tasks when they can communicate with each other to understand each other's strengths and weaknesses [7, 65]. Third, when tasks have subjective or unclear guidelines, deliberation can help workers converge on decisions [5, 6, 58]. Finally, collaborative environments help workers quickly divide tasks on the fly when it is difficult to automatically divide a job into microtasks [34, 43]. These systems, which rely on close worker interaction, highlight the value of collaborative crowdsourcing.

3 MOTIVATING GROUPS WITH FAIR PAYMENTS

Prior work has shown that workers on Mechanical Turk are primarily motivated by monetary rewards [54], and the impacts of various payments are well understood for individual work [18]. However, little is known about motivating workers through pay when their work is collaborative. The simplest payment method is to pay all workers the same amount, but equal payment does not recognize differences in skill or effort between the workers in the group. This shortfall may lead to a significant problem in worker motivation, as the best-performing workers could feel undervalued for their work. In this section, we formalize this idea with the framework of equity theory, propose two theoretically fair payment methods, describe how to measure workers' perceptions of fairness, and compare these concepts with payment methods in existing collaborative tasks.

3.1 Worker Motivation and Equity Theory

Equity theory [1] states that humans compare themselves to other people to decide whether they are being treated fairly. Humans believe that their outputs are equitable when

$$\frac{O_{self}}{I_{self}} = \frac{O_{other}}{I_{other}},$$

where I is one person's perceived input and O is their output. In other words, this relationship states that somebody that puts in twice as much work as their colleague should be rewarded twice as much. These outputs typically refer to some type of tangible reward, such as wages or bonuses. However, the inputs are not clearly defined. Depending on the situation, the inputs could be related to the amount of time spent working, the quantity of work done, or the quality of the work.

When workers do not believe that their outputs are equitable, they change their inputs to fix the discrepancy. In other words, overpaid workers will put in more effort, and underpaid workers will put in less effort. Workers might even quit their work in response to extremely unfair outcomes. It is crucial to ensure that workers do not feel underpaid, compared with other members of the group, to keep them motivated in collaborative work.

To make judgements about equity, workers must be able to see others' inputs. In microtask crowdsourcing, the availability of this information depends on the type of collaboration in their work. In tasks where workers have no knowledge of each other, they cannot compare inputs. However, when workers can *see others' work* and are *aware of others*, they can get a sense of the range of inputs that other workers are providing, giving them an approximate point of comparison. When workers can *identify others' work*, they can also make specific judgements about individual group members. These extra pieces of information help workers to judge whether their payments are equitable in collaborative work.

Workers also need to have access to others' outputs (payments) to make equity comparisons. This information is much more readily accessible than others' inputs. Workers often post details about their wages on public forums, such as Reddit's r/mturk or TurkerNation, or on task reviewing websites, such as Turkopticon [22] or TurkerView. Many workers also rely on personal connections, and it is common for them to discuss wages [63]. These channels can give workers an idea of the payment range for a task.

Additionally, requesters can make this payment information transparent, allowing workers to see others' exact rewards. Several authors have suggested that this added transparency would be beneficial. Martin et al. [44] concluded that additional market transparency would help workers focus on their tasks by eliminating "work to make Turking work". These impacts are magnified by the global nature of crowdwork [45]. Fieseler et al. [11] also advocated for increased transparency about workers' payments. They posited that this information would combat feelings that requesters are being deceptive about their workers' pay, making workers more loyal and improving trust and intrinsic motivation. Payment transparency could also help workers cope with unclear instructions by helping them recognize work that requesters marked as high- or low-quality. Overall, making payment information available would improve relations between workers and requesters, benefiting both parties.

Equity theory's predictions have been tested in a number of other settings. First, they have been validated extensively in laboratory studies. Mowday's review of this experimental work [51] found supporting evidence that overpaying leads to higher effort and underpaying to lower effort. Harder [16] also found support using data from professional baseball and basketball. His analysis showed that overpaid athletes performed better and acted more cooperatively, while underpaid athletes performed slightly worse and made more selfish plays. At a group level, position- and outcome-based rewards have been correlated with employee satisfaction and productivity [57],

and data from firms in Belgium and Sweden shows a relationship between unequal wages and productivity [17, 32]. We are not aware of prior work testing for these effects in crowd work.

3.2 Fair Payments

In this paper, we focus on a specific set of payment systems. We suppose that a requester posts a task where a group of workers earns a collective payment together. This payment could be fixed, as in many existing tasks, or it could include a performance-based bonus for the group. Then, the challenge of this system is to divide the group's payments among the individual workers.

The most basic payment method is to simply pay all workers equally. This method is the default in micro-task crowdsourcing: usually, workers received a fixed, pre-determined payment for submitting a task. However, equal payments do not recognize varying levels of skill and effort between workers in the group. Thus, we use equal payments as our control, and we propose two alternative group payment methods based on concepts from the literature.

The first alternative is to pay workers according to equity theory. In order to ensure that each equity judgment is satisfied, the ratio of each worker's output to input must be equal. This requirement means that the payment for worker i should be

$$O_i = c \cdot I_i,$$

where c is the amount of pay per unit of work. We note that there is still some subjectivity in this definition, as the input I could depend on several different metrics, such as work quality or quantity, or time spent on the task.

Another type of theoretically fair payment comes from the field of cooperative game theory [4]. A transferable utility cooperative game consists of a set of players N and a characteristic function $v(C)$ which describes the amount of reward that every possible subset of the players could earn by working together in a coalition C. There are numerous ways to divide the rewards between the players so as to satisfy different properties. One well-studied reward division is the *Shapley value*, which is focused on splitting the rewards fairly. The Shapley value for player i is

$$\phi_i = \sum_{C \subseteq N \setminus \{i\}} \frac{|C|!(|N| - |C| - 1)!}{|N|!} \left(v(C \cup \{i\}) - v(C)\right).$$

Intuitively, this is the average amount of value a player contributes when they join the group. This reward division satisfies four fairness axioms. It allocates the entire group's reward (efficiency), gives equal rewards to players that contribute the same amount (symmetry), gives no reward to players that contribute nothing (null players), and adds the rewards when combining two characteristic functions (additivity). We note that cooperative game theory also prescribes other reward divisions, such as the *core*, which focus on stability: they ensure that no rational player wants to leave their group. In this paper, we choose to focus on axiomatically fair rewards rather than stable rewards.

It is important to note that these theoretical methods cannot be applied to all types of work: both of them require a clear definition of workers' inputs. In some crowdsourcing tasks, there are no straightforward ways to compare workers. One example is in deliberation tasks, where describing an individual worker's contributions would require a deep understanding of the deliberation process. In this type of work, an alternative method for payment division is to ask workers how valuable their group members are. Algorithms for combining workers' subjective reports have been studied in the social choice literature [12]. However, these methods must recognize workers' conscious or unconscious biases toward themselves [55, 61] and stop workers from colluding with each other to increase their payments. We choose to leave these worker-determined payments for future work.

3.3 Measuring Perceptions of Fairness

In order to evaluate these theoretically fair payments, we need a method for measuring workers' perceptions of fairness. One way to compare group payments is to explicitly ask workers whether their payments are fair. Organizational justice is a construct that measures employees' perceptions of fairness in a workplace. Colquitt [8] summarized this literature by describing four different components of justice and a set of questions designed to measure each of these components. One of these components is distributive justice, which specifically focuses on the fairness of workers' outcomes. Colquitt showed that distributive justice is correlated with satisfaction: workers tend to be most satisfied with their outcomes when they feel that the distribution is equitable.

However, humans are not perfect at recognizing fairness: in fact, they are often significantly biased toward themselves [48]. There are multiple reasons for this effect. One reason is that people believe that their work is more valuable because they remember more facts about their own work than their colleagues. Another reason is that people may react more strongly to being underpaid than to being overpaid. Recognizing these biases is central to understanding the whole picture of workers' fairness perceptions.

3.4 Payments in Existing Tasks

Existing collaborative tasks have used a variety of payment systems. In Legion [35], workers were paid bonuses in proportion to a "power score" based on their agreement with the group. In Scribe [33], workers' audio transcriptions were combined using a sequence alignment algorithm, and they were paid bonuses based on the number of words that matched the final, aligned transcript. Kaspar et al. [24] had a user act as an "oracle", rating the quality of each workers' image segmentations, and paid bonuses according to these quality ratings. These payment systems are ad-hoc, and some can be rather opaque: it is difficult to workers to understand how their payments relate to their work quality, making it hard for them to make equity judgements. One final example is the manager-led teams in DreamTeam [65], where workers were paid bonuses for acting as the team's manager. This task is an example where asymmetric worker roles are paid different bonuses, according to the difficulty or value of the roles.

4 STUDY 1: PERFORMANCE-BASED BONUSES

In the previous section, we defined proportional payments and Shapley values, and we showed that these payments should be perceived as being more fair and should elicit more worker effort than equal payments. We performed a crowdsourced study to examine whether these effects can be observed in a real collaborative task. Specifically, this study attempts to answer three questions:

- **Question 1**: Do workers perceive proportional and Shapley value payments as being more fair than equal payments?
- **Question 2**: Are workers' fairness perceptions biased toward themselves?
- **Question 3:** Do workers put in more effort when they are paid fairly?

4.1 Method

To answer our three questions, we had workers complete a collaborative audio transcription task. We split performance-based bonuses between groups of workers using various bonus divisions, and we evaluated workers' fairness perceptions and performance levels based on these payment methods.

4.1.1 Participants. We hired participants from Mechanical Turk. We posted HITs with the title "Transcribe audio with a team of workers" and offered a base payment of $1.75. In the HIT instructions, we estimated that the HIT would take approximately 25 minutes, and we stated that workers

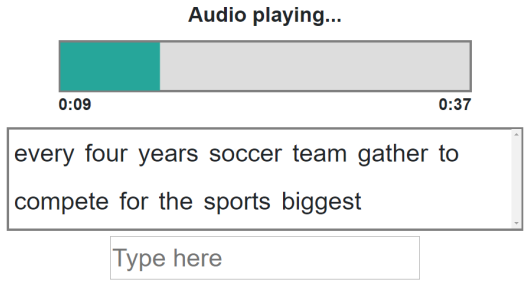

Fig. 1. The audio transcription interface. Workers listened to short audio clips and typed the words they heard in real time. Each audio clip ended with 7 seconds of silence to allow workers to finish typing.

would receive a performance-based bonus with a typical value of $1. We required workers to have at least 1000 approved HITs with a 95% or higher approval rate.

4.1.2 Groups. After workers accepted the HIT, we placed them into a 'virtual' group with two previous participants. We selected these group members by drawing randomly from the pool of workers that had finished the experiment. We ensured that workers could only be selected twice. We initialized this pool of workers with participants from a pilot study. We also informed workers that their data may be re-used to serve as coworkers in future batches of HITs. It was clear to the workers that they were not working together in real time.

4.1.3 Task. For our experimental task, we used a real-time audio transcription task based on Scribe [33]. Workers were not allowed to pause or replay the audio, as if the transcript was required in real time. This task is suitable for several reasons. First, it is a difficult task, and workers need to focus to produce high-quality transcripts. Second, it is impossible for a single worker to produce a perfect transcript, motivating the need for multiple workers to complete the same task. Third, it is easy to learn, as many workers are familiar with regular audio transcription tasks. Finally, it is realistic: this interface could be used for a real-time captioning task. Our transcription interface is shown in Figure 1.

During the experiment, workers were *aware of others* and could *see* and *identify others' work*, but could not *freely interact*. We chose this combination of features intentionally. In order to make equity judgements, workers must be *aware of others* and *see others' work*; without this information, they cannot compare their inputs with each other. We also let workers *identify others' work* to allow them to make equity judgements about specific teammates, rather than the group as a whole. We chose not to allow workers to *freely interact*. Prior work has shown that personality differences have a strong influence on workers' satisfaction [39], and we attempted to limit this effect by avoiding open communication. We discuss how these choices affect our results in Section 6.3.

4.1.4 Procedure. In the experiment, workers first filled out a consent form and completed an interactive tutorial about the interface. Then, they transcribed 14 short audio clips that we manually selected from podcast episodes.[1] We used podcasts for our audio clips because there were high-quality transcripts available as a source of ground truth. The audio clips varied from 21 to 31 seconds with a median length of 28 seconds. We added an additional 7 seconds of silence to the end of each clip to allow workers to finish typing. We processed each word that workers typed by removing all punctuation and converting the text to lowercase. Then, at the end of each audio clip,

[1]We used podcasts from http://freakonomics.com/.

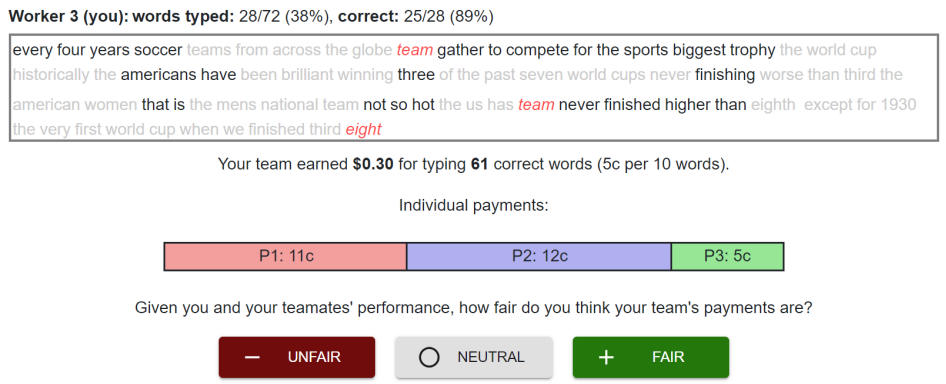

Fig. 2. The bonus payment screen, showing an example of one worker's transcript. Workers saw the full text that each member of the group typed, how these transcripts compare to the ground truth, and the exact bonus that each worker received. (Workers could see all three group members' transcripts; to save space, we only show one here.)

we compared workers' transcripts to the ground truth with a word-level Myers diff [52], which allowed us to check whether workers typed each word correctly.

4.1.5 Bonuses. After each audio clip, we showed workers how well each member of their virtual group performed. We summarized each worker's performance by displaying both the number of words typed and the number of correct words. We also showed workers the full diff output, with correct words in black, incorrect words in red, and untyped words in gray, allowing them to interpret these results. Next, we counted the number of words in the ground truth transcript that were correctly typed by at least one worker. We calculated a total bonus payment of 5 cents for every 10 words that the group collectively typed correctly. We selected this bonus scale so a typical group would earn a bonus of 20 to 30 cents per round. The payment screen is shown in Figure 2.

After calculating the group's bonus, we divided it between the three workers. We placed groups into one of four experimental conditions:

- Equal: We gave each worker one third of the group's bonus. This method is the control, as it is similar to the default of paying a fixed HIT reward.
- Proportional: We counted the number of words that each worker typed correctly. Then, we gave each worker a bonus proportional to the number of correct words that they typed. This method is fair according to equity theory.
- Shapley: We computed the bonuses that each of the 8 subsets of the workers would have earned. Then, we paid workers with the Shapley values, using these bonuses as the characteristic function. This method is fair according to cooperative game theory.
- Unfair: As a manipulation check, we gave 50% of the bonus to the worker that typed the smallest number of words correctly, and we gave 25% of the bonus to the other two workers.

In all four cases, we rounded bonuses down to the nearest cent. We displayed the transcripts and bonuses to workers in a payment screen at the end of each round, shown in Figure 2. Finally, we asked workers to rate the division of bonuses as 'Fair', 'Neutral', or 'Unfair' before proceeding to the next audio clip.

4.1.6 Post-Study. After transcribing all 14 audio clips, workers filled out a post-study survey. In the survey, we asked five 5-point Likert scale questions about the bonus payments. We adapted

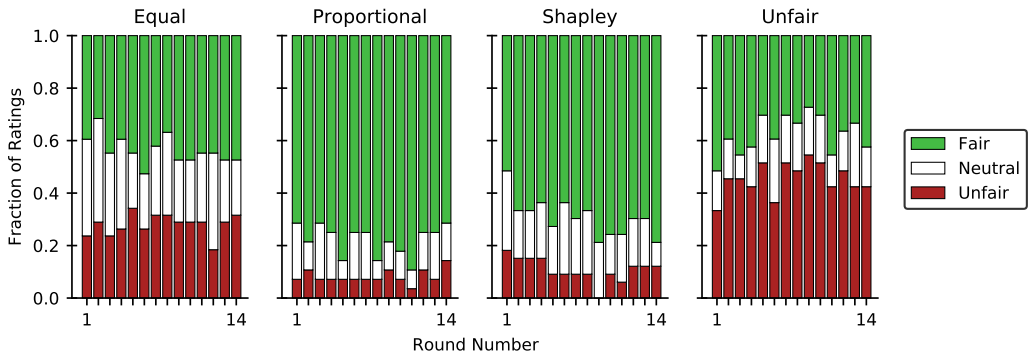

Fig. 3. Workers' fairness ratings for each round of the experiment. Workers in the Proportional and Shapley conditions rated their payments as being more fair than workers in the Equal or Unfair conditions.

these questions from Colquitt's distributive justice and satisfaction measures [8]. Specifically, we asked workers whether their payments were appropriate, justified, acceptable, and satisfying, and whether the bonuses reflected the effort they put into the task. We also asked workers about their demographics, how they selected their fairness ratings, whether they enjoyed the task, and their feelings about working in a group with other workers. Lastly, after workers submitted the HIT, we granted bonuses to all three of the group members – both the participant and their two virtual coworkers.

4.2 Results

A total of 132 workers completed the HIT. We removed 2 workers that typed 0 words in the first round of the task. The number of workers in each condition varied from 28 to 38 workers; we confirmed that these conditions were not significantly unbalanced with a chi-squared test ($p = 0.65$). Workers typed an average of 29.23 words per round ($\sigma = 10.55$), with 24.48 of these words being marked as correct ($\sigma = 9.87$). Overall, the median worker spent 22.3 minutes on the HIT and earned a bonus of 96.5 cents, resulting in a wage of \$7.30/hour[2].

4.2.1 Fairness Ratings. Each worker submitted one fairness rating for each of the 14 rounds in the main experiment. These ratings are plotted in Figure 3. This plot shows that workers are most likely to rate their payments as fair in the Proportional and Shapley conditions. To confirm these differences, we fit a proportional odds model to these ratings using the workers' conditions as a factor. This model showed that ratings in the Equal condition were significantly more negative than the Proportional ($p < 0.001$) and Shapley conditions ($p = 0.002$), but not significantly different from the Unfair condition. Thus, the answer to our first research question is yes: workers do recognize theoretically fair payments as being more fair than equal payments.

4.2.2 Worker Bias. We also investigated the amount of bias in workers' fairness ratings. To do this, we split workers' ratings across all rounds into three groups: whether they were the best, the middle, or the worst worker in their group for each round. The distribution of ratings for each condition and group position is shown in Figure 4. This plot suggests that workers' perceptions of fairness change based on their abilities.

[2]Note that workers also earned up to two additional bonuses if their transcripts were reused in a future team.

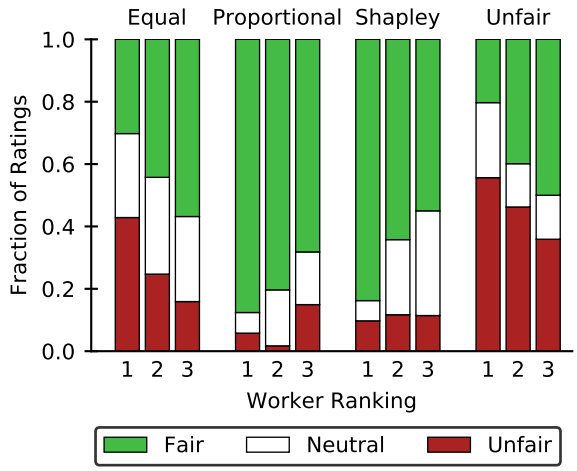

Fig. 4. Fairness ratings for each condition, split by workers' ranking in the group. The best worker in each round has a ranking of 1, and the worst has a ranking of 3.

We confirmed these biases by adding a measure of the workers' relative skill levels to our proportional odds model. For each round, we calculated the skill difference between the participant and their two teammates as

$$\begin{aligned}\textsc{Skill Difference} = 2 \cdot \textsc{Words Correct}_{worker} \\ - \textsc{Words Correct}_{\text{coworker 1}} \\ - \textsc{Words Correct}_{\text{coworker 2}}.\end{aligned}$$

This quantity is positive when the participant types more correct words and negative when they type less correct words than their coworkers. After adding this factor to the model, the results showed that Skill Difference had a negative effect in the Equal ($p = 0.006$) and Unfair ($p < 0.001$) conditions: workers with more skill than their coworkers thought that these payments were less fair. On the other hand, it had a positive effect in the Shapley condition ($p < 0.001$), where workers felt their pay was more fair when they were more skilled. Finally, Skill Difference had no significant effect in the Proportional condition.

4.2.3 Justice Ratings. Workers' answers to the five post-survey Likert scale questions had a high level of internal reliability (Cronbach's $\alpha = 0.92$). We aggregated these answers into a single justice score for each participant by taking the average of the five answers. The resulting justice scores are shown in Figure 5. This boxplot shows that the score distributions are not the same: workers in the Proportional and Shapley conditions never give very low scores. However, the median scores in the Equal, Proportional, and Shapley conditions are quite similar.

We used non-parametric statistics to analyze these ratings.[3] A Kruskal-Wallis test revealed that the condition had a significant effect on the justice scores: $H(3) = 18.42$, $p < 0.001$. We performed post-hoc Mann-Whitney tests with a Holm-Bonferroni correction and found significant differences between the Proportional and Unfair conditions ($p < 0.001$) and between the Shapley and Unfair conditions ($p = 0.01$). All other comparisons were not significant. This analysis shows that workers responded more favourably to the theoretically fair payments than to the unfair payments.

[3]We first fit a one-way ANOVA model to the ratings, but a Shapiro-Wilk test showed that the residuals were not normally distributed ($p < 0.05$).

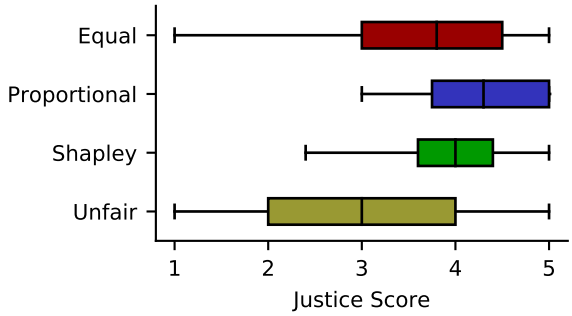

Fig. 5. A boxplot of workers' justice scores in each of the conditions. Workers had higher justice scores in the Proportional and Shapley conditions than in the Unfair condition; no other comparisons were significant.

Table 2. Workers' average change in performance between the first and the last round. Workers in the Proportional and Shapley conditions improved more than in the Equal condition, but no comparisons were significant.

Condition	Words Typed	Words Correct
Equal	M=4.03, σ = 5.10	M=4.97, σ = 5.03
Proportional	M=5.43, σ = 6.33	M=5.93, σ = 5.14
Shapley	M=6.82, σ = 6.54	M=7.30, σ = 6.75
Unfair	M=3.87, σ = 6.16	M=4.68, σ = 6.95

The differences between workers' justice scores in each condition were quite small. This effect contrasts with the fairness rating analysis, where the differences between the four conditions were more clear. This effect may be caused by the timing of these questions. In the post-survey, workers may have considered their bonus payment for the entire experiment and answered whether it is fair, compared to typical Mechanical Turk wages. As our task paid more than the median wage on Mechanical Turk—approximately $2 per hour [15]— workers may have tended to answer more positively than expected. Alternatively, workers may have been hesitant to select the "extreme" answer of 1 for the justice questions.

4.2.4 Effort. We recorded two performance metrics in each round: the number of words each worker typed and the number counted as correct. These metrics are affected by many factors, including the length and difficulty of the audio clips, as well as the workers' skill and effort levels. We chose to consider each worker's change in performance between the first and last rounds. Comparing these changes between conditions allows us to isolate the workers' learning rates and effort levels. The average changes are shown in Table 2. These values suggest that there may be a small difference in performance improvements between the conditions, with workers improving by 1 to 3 more words in the Proportional and Shapley conditions.

To analyze these differences, we fit two binomial regression models: one to Words Typed and another to Words Correct. In both of the models, we fit the workers' final round performance, using their condition and first round performance as factors. For both models, we found a main effect of first round performance ($p < 0.001$), but no main effects of condition or interaction effects. In other words, we could not detect any significant differences in performance changes between the

conditions. To validate this result, we compare our results to previous work on bonus payments for crowdsourcing tasks. Ho et al. [18] found that workers corrected 1 additional error out of 12 when they were paid with appropriate bonuses. This improvement—an increase of less than 10%—was only detected with large samples of up to 1000 workers. We suggest that studying workers' effort requires more accurate measurements of their baseline skill and tasks with less variation in their individual performance.

4.2.5 Survey Responses. Workers had a variety of explanations for their fairness ratings. Many workers mentioned making direct comparisons between the number of words or accuracy of their group members. Others explicitly referred to the effort that they put into the task. Another common theme was the difficulty of the task: several workers were surprised that real-time audio transcription was so difficult. In particular, workers that thought they performed poorly often said that they were happy to get any bonus at all. We note that these feelings might affect workers' opinions about their payments: if they believe that they did poorly in the task, then they might be less critical of their bonuses.

Workers had diverse opinions about how enjoyable the task was. Negative comments tended to mention how frustrating, difficult, tedious, or weird the task was. Positive comments described the task as fun, challenging, or different from usual HITs. Workers were also split about the competitive aspect of the task: some workers enjoyed the competition, while others thought it was stressful to compare themselves against their group.

Many workers were positive about working in a group. They described it as being motivating and fun, while helping them to earn larger bonuses. They also mentioned that having multiple workers do the same task can make for useful feedback, allowing them to learn from each other. Even some workers that performed poorly enjoyed working with a group: they thought that their more skilled teammates helped them complete a task that they could not do alone. The negative comments argued that teamwork was more stressful, and some workers disliked the idea of relying on others.

The worst workers in the unfair condition rarely mentioned that they were overpaid. Some commented on the difficulty of the task, saying that it was hard to listen to the audio while also typing and spell-checking; one suggested that we slow down the audio. Others talked about their performance, acknowledging that they were much worse than their group, and one said that they would prefer groups closer to their skill level. The workers most critical of the unfair bonus system were the ones who performed the worst, though only in a small number of rounds.

5 STUDY 2: EXTERNAL RATINGS

In our first study, we examined how workers respond to different payment methods for collaborative work. Now, in our second study, we used an independent group of workers to review the bonus payments from the first study. We used this second set of opinions to look for additional biases in the original workers' fairness ratings.

5.1 Method

5.1.1 Participants: We hired participants from Mechanical Turk by posting HITs with the title "Review work done by other workers". We offered a HIT payment of $1.50 with no bonus. The HIT instructions gave a time estimate of 12 minutes. We required workers to have at least 1000 approved HITs with a 95% or higher approval rate. We also ensured that workers who completed the first experiment could not participate.

5.1.2 Task: In the second study, workers did not complete any audio transcriptions. Instead, we showed them transcripts from previous groups of workers and asked them to rate how fair the

bonus payments were. We used the same bonus payment screen except for minor modifications to the text (e.g., we changed "you and your teammates" to "the workers").

5.1.3 Procedure: After workers accepted the HIT, they accepted a consent form and completed a tutorial. In the tutorial, we explained the real-time audio transcription task so that workers understood the difficulty of the work. We also showed workers the bonus payment screen and asked comprehension questions about the transcript displays and bonus divisions. Then, workers were shown a total of 16 rounds from the audio transcription tasks. For each worker, we picked 3 random rounds from each of the 4 payment divisions. We also selected 1 fixed round for each payment division to show to every worker. These 16 rounds were randomly ordered. For each round, they clicked on one of three buttons, labelled "Fair", "Neutral", and "Unfair". As an attention check, we randomized the positions of the "Fair" and "Unfair" buttons in every round.

At the end of the study, workers filled out a post-study survey. We asked about their demographics, their reasoning for their fairness ratings, and whether they would like to rate or be rated by other workers in crowdsourcing tasks. Finally, workers submitted the HIT.

5.2 Results

A total of 79 workers completed the HIT. We removed 16 workers that averaged less than 5 seconds per round, leaving 63 workers. After this filtering step, we did not find any workers that clearly ignored the task instructions. For clarity, in this section we refer to the new participants as the external raters, and we refer to the participants from Study 1 as the original workers.

5.2.1 Fairness Ratings: Workers submitted a total of 1008 ratings: 756 on the randomly selected rounds and 252 on the fixed rounds. We found that the original workers' ratings on the fixed rounds were not representative of typical ratings in each condition, so we chose to focus only on the randomly selected rounds. The aggregates of these ratings are shown in Figure 6. This plot suggests that raters were generally more critical than the original workers, rating "Unfair" more often. This effect is strongest for the Equal and Unfair payments.

We first analyzed the external raters' ratings alone for each condition. To do this, we fit a proportional odds model to the ratings using only the payment method as a factor. This model shows significant differences between the Equal payments and each of the other three payment methods (all $p < 0.001$). Post-hoc tests with a Holm-Bonferroni correction showed significant differences between each of the conditions ($p = 0.002$ for Proportional − Shapley; all other comparisons $p < 0.001$). The directions of these post-hoc tests show that the Proportional payments were rated as the most fair, followed by Shapley payments, then Equal payments, with Unfair payments being rated as the least fair.

We also compared the original workers' fairness ratings with the external raters' to check for differences between these two groups of workers. For each condition, we performed a paired Wilcoxon signed-rank test between the two sets of ratings. These tests showed that the external raters found the payments less fair than the workers for the Equal ($p = 0.004$), Shapley ($p < 0.001$), and Unfair ($p < 0.001$) conditions. We found no significant difference in the Proportional condition ($p = 0.22$), suggesting that the external raters and the workers shared similar opinions about these payments.

We suggest several possible reasons for the differences in ratings between the two groups of workers. First, the original workers only saw one type of payment, while the external raters saw all four types. Workers may be more critical of Equal pay if they are aware of the other, theoretically fair payments. Second, external raters are not biased in the same ways that the original workers are. It is easier for raters to honestly judge whether a payment is fair because they do not benefit from the payments.

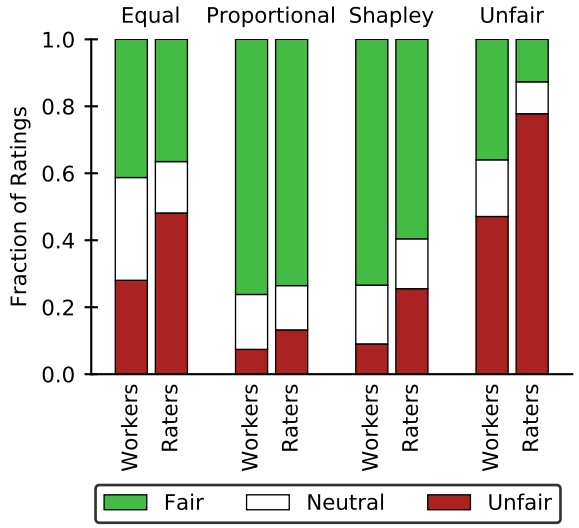

Fig. 6. Fairness ratings from the original workers in study 1 (left), compared with the external raters in study 2 (right). Raters were more critical of the EQUAL, SHAPLEY, and UNFAIR payments.

5.2.2 Survey Responses: The external raters judged fairness using similar criteria to workers in the first study. Most of the responses mentioned comparing the group members' numbers of words typed, correct words, or accuracy. A few raters were more interested in effort, and looked more carefully at the words that the group members typed in order to gauge how hard they were working. Some workers explicitly referred to their overall wages on Mechanical Turk, with one worker citing "hourly wage... how much I work to eat."

The majority of the workers were positive about the idea of rating each others' work, as long as they were paid to do it. Most workers were also happy to have their work judged by others. One worker pointed out that this is already close to their job: requesters can judge every HIT that they submit. However, several participants disagreed, saying that this felt invasive and that it would be hard to trust the raters. Finally, one response said that it would be stressful having to worry about performance ratings on top of already low pay.

6 DISCUSSION

In this paper, we studied how crowd workers are motivated by different payment divisions for group-based work. We identified two theoretically fair payments, motivated by equity theory and cooperative game theory, and discussed their relationship with crowd workers' motivation. Then, results from our first study show that workers who were paid theoretically fair bonuses—that is, proportional to quality of their work or calculated with the Shapley values—reported their payments as being more fair than equal bonuses. Furthermore, our second study showed that this effect is even stronger for external raters that were not involved in the tasks. Workers were mostly positive about tasks that involve working with or evaluating other workers. Finally, our performance metrics suggest that workers might exert slightly more effort when they are paid with these fair bonus divisions, but we do not have conclusive evidence of this effect. In this section, we discuss the implications of our findings for future collaborative crowd work.

6.1 The Impacts of Payments and Transparency

Our literature review showed that collaborative tasks with close interactions between workers can be used to solve complex problems. By allowing these close interactions, groups of workers can combine their skills in creative writing or cognitive tasks, work around subjective instructions, or divide work on the fly. However, in these tasks, it is essential to ensure that workers feel that they are paid equitably, and our experimental results showed that workers are receptive to fair and unfair payments. In order to keep workers motivated and satisfied with their rewards, it is crucial to pay workers relative to their contributions to the group. Further, our results on worker biases revealed that the most skilled workers are also the most negatively affected by unfair pay, giving requesters even more reason to use fair payments.

Fieseler et al. [11] posited that treating workers fairly and transparently is not simply a question of ethics. They proposed several potential benefits for both workers and requesters that could be realized through features on crowdsourcing platforms. Allowing communication between workers can decrease feelings of isolation, help workers set time commitments and effort levels, and clarify task descriptions. Additionally, payment transparency helps to mitigate feelings that requesters are lying or deceiving workers. Together, these effects lead to more committed workers with increased trust, satisfaction, and intrinsic motivation. We argue that collaborative tasks are an opportunity for requesters to reap these benefits now. Rather than relying on the platform to take action, requesters can implement tasks with explicit collaboration and public payment information. As long as requesters select equitable payments, these tasks are an excellent chance to build trust and reputation with workers, and ultimately to produce better results.

We reiterate that it is impossible to keep payment information fully hidden. Crowd workers have a basic social need for communication [13]; when they are not provided with communication channels, they seek to reproduce these channels in public forums and private companionships. These external communication lines give workers a way to exchange payment information, and these discussions can often be more speculative than truthful. On top of requesters' moral duty to treat workers fairly, we also believe it is in requesters' best interests to communicate with workers on public platforms like Turkopticon and TurkerView, or even publicize payment information themselves.

There is also an opportunity here for crowdsourcing platforms to make an impact. Mechanical Turk hides most information about its workers, and requesters—particularly inexperienced ones—may interpret this anonymity as a signal that workers do not communicate with each other. This lack of information can encourage opportunistic or exploitative behaviour from requesters [11]; in fact, even well-intentioned requesters cannot correct their actions unless they know that their workers are unhappy. Platforms can combat this behaviour by improving transparency on their marketplace: for instance, by displaying requesters' historical wages on the workers' interface. Although many workers already rely on external tools that provide this information, building these features into the platform would send a clear signal to requesters that they should be conscious about treating workers equitably.

A substantial number of workers from our studies were intrinsically motivated by working with others, describing the teamwork as enjoyable, motivating, and fun. For these workers, it would be useful to provide a consistent source of collaborative work. Gray et al. [13] proposed splitting crowd work into two separate streams, with one stream permitting collaboration in tasks that do not require independent responses. We suggest that this idea can be taken another step further. Rather than simply allowing workers to communicate, this stream of work can be designed to leverage the benefits that workers and requesters receive from transparent teamwork. This split

would also be helpful for workers that do not enjoy group work, and would prefer to continue working alone.

6.2 Fair Payment and Effort

In our main experiment, we did not find conclusive evidence that workers exert more effort when they are paid using theoretically fair methods. It is possible that there truly is no effect. Mason and Watts [46] found that work quality was not affected by payments due to an anchoring effect. Ho et al. [18] also suggested that performance-based payments may not affect worker effort if the bonuses are too small, relative to the task's overall pay, or if the task is not effort-responsive. We used a relatively small bonus compared to our base payment so that even the lower-performing workers could earn close to minimum wage in our experiment. However, there are several other possible explanations for our results.

First, real-time audio transcription tasks are not perfectly suited for measuring a worker's skill and effort. Our metrics, which are related to typing speed and accuracy, have a large amount of variance between audio clips. Future studies on this topic should consider tasks where the quality of workers' output is more consistent, and should more carefully measure workers' initial skill levels during training rounds or qualification tasks.

The other reasons are factors that could affect workers' motivation and actions. We paid workers close to minimum wage, which is substantially higher than a typical task on Mechanical Turk [15]. We also told workers that they would be paid bonuses. Knowledge of a bonus might reduce workers' fear of having their work rejected, as bonuses on Mechanical Turk can only be paid after approving a HIT. Without this knowledge, workers might have worked harder to ensure their work is accepted, even if their bonuses are not motivating.

Finally, many workers mentioned that they found the task fun, interesting, and different. Workers that are intrinsically motivated might work hard regardless of their group's bonuses. Tedious, uninteresting tasks such as Yin et al.'s button-clicking task [62] would help to isolate the effects of bonuses on workers' effort. Longer tasks, taking an hour or more, would also help to capture these effects.

6.3 Generalizing to Other Tasks

In our experimental task, workers were *aware of others* and could *see* and *identify others' work*, but could not *freely interact*. How would our results change if we used a task from a different category?

It would not be possible to perform our experiment if workers were not *aware of others* or could not *see others' work*. In this case, workers cannot make immediate equity judgements: either they cannot see others' inputs, or they do not know that other workers created them. In these types of tasks, though, it may still be possible for workers to understand the relative quality of their work: they often discuss their work on public forums and through private communications. In the long term, if these tasks do not pay fairly, we believe that skilled workers would still become frustrated with their payments.

Workers could still make equity judgements if they could not *identify others' work*. Even without specific details about each of their group members, workers could still understand whether they were more or less skilled than their teammates. This information would give them a sense of the proportion of the group's reward that they deserve. Thus, while workers would not be able to rate whether their teammates were being paid fairly, they would still be able to rate their own payments. We believe that our results would be qualitatively similar in these types of tasks.

The ability to *freely interact* could have a number of different effects on our results. Giving workers an open communication line would help them understand how much work their teammates are putting into their tasks. In some cases, this might help workers build common ground and trust with

each other, making them more supportive of equal payments regardless of the group's skill levels. In others, it might reveal differences in the workers' effort levels, making the better workers feel that they deserve a larger share of the reward. These effects are likely dependent on the workers' personalities and the complexity of the task, and understanding how these factors influence fairness perceptions is an important avenue for future work.

6.4 Limitations

In our main experiment, we simulated a group environment by comparing workers' transcripts against previous participants. This style of task is similar to existing crowdsourcing workflows, but it is quite different from tasks with real-time group interactions. Working with a group in real time may be more motivating, but it could make workers more frustrated or anxious as they are forced to work at the group's pace. More work is required to understand the impacts of real-time interaction.

We chose to focus our experiments on groups with three workers, as these small groups are common in existing collaborative tasks. Increasing the size of the groups could affect our results in several ways. Larger groups might impact workers' equity judgements, as it may be more difficult to make equity judgements when there are more possible choices of "other". This increased difficulty might make equal payments more agreeable. Group size can also impact worker effort in collaborative crowdsourcing tasks: prior work has shown that larger groups suffer from increased social loafing [43].

Finally, we did not control for the location of the workers in our experiment. The majority of workers on Mechanical Turk are located in the United States, but an appreciable number live in other countries, with the largest group being from India [13]. It is possible that there are significant cultural differences between these worker populations that we have not studied here.

6.5 Broader Impacts

Crowdsourcing platforms incentivize low pay, with workers on Mechanical Turk making a median wage under US$2 per hour [15]. Finding new, difficult problems that workers can solve together could have the unfortunate consequence of attracting more low-paying requesters to the system. However, we believe that tasks with explicit teamwork are beneficial to the workers. As we described above, having workers cooperate can give them more information about their work, making it easier for them to avoid returning HITs or being rejected for misunderstanding a task – two of the biggest impacts on their hourly wage [15].

Our proposed payments may appear to be in conflict with minimum wage standards. When one worker does not produce any useful input, both the proportional payments and Shapley values give no payment. This point could be an issue: sometimes, workers cannot complete their work due to factors out of their control, such as poor UI or unclear instructions. However, both payments can easily adapt around this issue. For proportional payments, each worker's input can combine the amount of work they did with the amount of time they spent on the work, ensuring a minimum wage. Similarly, the Shapley value can also be modified by relaxing the null player axiom: egalitarian Shapley values [23] can describe any convex combination of equal pay and the Shapley value. These adjustments allow the principles of equity theory to be applied while still ensuring an ethical minimum wage.

Lastly, performance-based payments also give lower payments to less productive group members. This low pay can be problematic: it can be stressful to less experienced workers or unfair to workers with disabilities. However, this issue is not unique to collaborative work. In non-collaborative tasks, even without performance-based bonuses, less experienced workers are more likely to spend more time on HITs or have their work rejected, leading to lower wages.

However, we believe that our payment systems can also provide some tools to make positive impacts for these workers. First, equity theory suggests that the payments should be proportional to workers' inputs, but does not specify what these inputs measure. They could include the effort that workers spent on the task, rewarding hard work regardless of its quality. It is an interesting challenge to ensure that these payments still incentivize high effort. Second, transparent collaboration can help workers learn by seeing high quality work. Working together closely with more skilled teammates can help inexperienced workers improve by understanding their mistakes. For new workers that do not have experienced friends to rely on for advice, collaborative tasks could make crowdsourcing more inclusive.

6.6 Future Work

Our research raises several interesting questions and we see three key directions for future work. The first is to study how workers rationalize about their inputs. When judging the equity of a payment, workers could base their decision on work quality, quantity, or time. While some might value polished work, others might appreciate effort or feel that they deserve pay just for "showing up". Workers may also put different weights on these factors depending on their skill level, experience with the task, and relationships with each other.

The second direction for future work is to adapt these group-based payments to tasks where work quality is difficult to measure and convey to other workers. For instance, the performance based payments in our experiment could not be calculated for a real, uncaptioned audio stream. To deal with this uncertainty, workers' outputs could be compared with an agreement score [35], with peer prediction algorithms [49], or by measuring how much weight an algorithm places on each worker's output [33]. In these cases, the reward calculations will likely appear opaque, and it is not obvious whether the workers will react favourably to these mechanisms. An alternative method is to pick bonuses based on workers' ratings of each others' work. Combining subjective ratings this way would be suitable for tasks like collaborative design, but finding ways to deal with collusion between workers is a significant challenge in this scenario. External, unbiased workers could also be asked to rate work quality; our participants confirmed that this would be a reasonable task.

Finally, a number of tasks involve collaboration between human workers and AI agents. Two examples in this area are Evorus [21], where chatbots suggest messages alongside human workers, and DreamTeam [65], where groups of workers are managed by Slack bots. In these tasks, workers could potentially feel that these AI agents are taking their work, lowering their pay. As human computation systems continue to incorporate more computational agents, it will be increasingly important to understand how workers' perceptions of equity and fairness change in these new types of tasks.

7 CONCLUSION

Allowing crowd workers to collaborate is a powerful tool, but motivating groups of workers is an entirely different challenge from ordinary crowdsourcing tasks. In this paper, we identified existing classes of collaborative crowdwork and proposed two theoretically fair methods for dividing payments between a group of workers using concepts from social psychology and cooperative game theory. We evaluated these payment systems on two groups: we asked the workers receiving these bonuses about their perceptions of fairness, and we compared these results against a group of unbiased raters. Our results show that workers recognize equal payments as less fair than either of these theoretically fair methods, but are biased toward payments that favour themselves. Understanding and implementing fair payment divisions is paramount to developing collaborative crowdsourcing systems for solving complex problems.

ACKNOWLEDGMENTS

Thanks to the Mechanical Turk workers for participating. We also thank Alex Williams, Mike Schaekermann, and four anonymous reviewers for their helpful comments on the paper. We acknowledge the support of the NSERC Discovery Grant (RGPIN-2015-0454), the NSERC CGS-M scholarship, and the Ontario Graduate Scholarship.

REFERENCES

[1] John Stacy Adams. 1965. Inequity In Social Exchange. In *Advances in Experimental Social Psychology*. Vol. 2. 267–299. https://doi.org/10.1016/S0065-2601(08)60108-2

[2] 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 Press, New York, New York, USA, 33. https://doi.org/10.1145/2047196.2047201

[3] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*. ACM Press, New York, New York, USA, 313. https://doi.org/10.1145/1866029.1866078

[4] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. 2011. Computational Aspects of Cooperative Game Theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 6 (oct 2011), 1–168. https://doi.org/10.2200/S00355ED1V01Y201107AIM016

[5] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, New York, New York, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[6] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 ACM annual conference on Human Factors in Computing Systems - CHI '19*.

[7] D. Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A. Hearst. 2015. Structuring Interactions for Large-Scale Synchronous Peer Learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, New York, New York, USA, 1139–1152. https://doi.org/10.1145/2675133.2675251

[8] Jason A. Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86, 3 (2001), 386–400. https://doi.org/10.1037//0021-9010.86.3.386

[9] J H Davis, P R Laughlin, and S S Komorita. 1976. The Social Psychology of Small Groups: Cooperative and Mixed-Motive Interaction. *Annual Review of Psychology* 27, 1 (1976), 501–541. https://doi.org/10.1146/annurev.ps.27.020176.002441 arXiv:https://doi.org/10.1146/annurev.ps.27.020176.002441

[10] Ryan Drapeau, Lydia B Chilton, and Daniel S Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)* (2016), 32–41.

[11] Christian Fieseler, Eliane Bucher, and Christian Pieter Hoffmann. 2017. Unfairness by Design? The Perceived Fairness of Digital Labor on Crowdworking Platforms. *Journal of Business Ethics* (jun 2017). https://doi.org/10.1007/s10551-017-3607-2

[12] Ya'akov (Kobi) Gal, Moshe Mash, Ariel D. Procaccia, and Yair Zick. 2016. Which Is the Fairest (Rent Division) of Them All?. In *Proceedings of the 2016 ACM Conference on Economics and Computation - EC '16*. ACM Press, New York, New York, USA, 67–84. https://doi.org/10.1145/2940716.2940724

[13] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The Crowd is a Collaborative Network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*. ACM Press, New York, New York, USA, 134–147. https://doi.org/10.1145/2818048.2819942

[14] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 2258–2270. https://doi.org/10.1145/2858036.2858364

[15] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–14. https://doi.org/10.1145/3173574.3174023 arXiv:1712.05796

[16] Joseph W. Harder. 1992. Play for Pay: Effects of Inequity in a Pay-for-Performance Context. *Administrative Science Quarterly* 37, 2 (1992), 321–335. http://www.jstor.org/stable/2393227

[17] Fredrik Heyman. 2005. Pay inequality and firm performance: evidence from matched employerâĂŞemployee data. *Applied Economics* 37, 11 (2005), 1313–1327. https://doi.org/10.1080/00036840500142101 arXiv:https://doi.org/10.1080/00036840500142101

[18] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*. ACM Press, New York, New York, USA, 419–429. https://doi.org/10.1145/2736277.2741102

[19] Shih-Wen Huang and Wai-Tat Fu. 2013. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 621. https://doi.org/10.1145/2470654.2470743

[20] Ting-Hao Kenneth Huang, Joseph Chee Chang, Saiganesh Swaminathan, and Jeffrey P. Bigham. 2017. Evorus: A Crowd-powered Conversational Assistant That Automates Itself Over Time. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17*. ACM Press, New York, New York, USA, 155–157. https://doi.org/10.1145/3131785.3131823

[21] Ting-Hao (Kenneth) Huang, Walter S. Lasecki, Amos Azaria, and Jeffrey P. Bigham. 2016. "Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*. 79–88.

[22] 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 Press, New York, New York, USA, 611. https://doi.org/10.1145/2470654.2470742

[23] Reinoud Joosten. 1996. *Dynamics, equilibria, and values*. Ph.D. Dissertation. Maastricht University.

[24] Alexandre Kaspar, Genevieve Patterson, Changil Kim, Yagiz Aksoy, Wojciech Matusik, and Mohamed Elgharib. 2018. Crowd-Guided Ensembles: How Can We Choreograph Crowd Workers for Video Segmentation?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–12. https://doi.org/10.1145/3173574.3173685

[25] Nicolas Kaufmann, Thimo Schulze, and Daniel Viet. 2011. More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*. ACM Press, New York, New York, USA.

[26] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P. Bigham, Jaime Teevan, Ece Kamar, and Walter S. Lasecki. 2017. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*. 89–97.

[27] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. 2017. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 233–245. https://doi.org/10.1145/2998181.2998196

[28] Joy O Kim and Andres Monroy-Hernandez. 2016. Storia: Summarizing Social Media Content based on Narrative Theory using Crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*. ACM Press, New York, New York, USA, 1016–1025. https://doi.org/10.1145/2818048.2820072

[29] 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 Press, New York, New York, USA, 43. https://doi.org/10.1145/2047196.2047202

[30] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, New York, New York, USA, 1003. https://doi.org/10.1145/2145204.2145354

[31] Anand Pramod Kulkarni, Matthew Can, and Björn Hartmann. 2011. Turkomatic: Automatic, Recursive Task and Workflow Design for Mechanical Turk. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*.

[32] Thierry Lallemand, Robert Plasman, and François Rycx. 2009. Wage structure and firm productivity in Belgium. In *The structure of wages: An international comparison*. University of Chicago Press, 179–215.

[33] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*. ACM Press, New York, New York, USA, 23. https://doi.org/10.1145/2380116.2380122

[34] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. 2015. Apparition: Crowdsourced User Interfaces that Come to Life as You Sketch Them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 1925–1934. https://doi.org/10.1145/2702123.2702565

[35] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology -*

UIST '11. ACM Press, New York, New York, USA, 23. https://doi.org/10.1145/2047196.2047200

[36] Jesse D. Lecy and Kate E. Beatty. 2012. Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis. *SSRN Electronic Journal* (2012). https://doi.org/10.2139/ssrn.1992601

[37] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. ACM Press, New York, New York, USA, 68. https://doi.org/10.1145/1837885.1837907

[38] Weichen Liu, Sijia Xiao, Jacob T. Browne, Ming Yang, and Steven P. Dow. 2018. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *ACM Transactions on Social Computing* 1, 1 (jan 2018), 1–26. https://doi.org/10.1145/3159649

[39] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. 2016. Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 260–273. https://doi.org/10.1145/2818048.2819979

[40] Ioanna Lykourentzou, Robert E Kraut, and Steven P Dow. 2017. Team Dating Leads to Better Online Ad Hoc Collaborations. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 2330–2343. https://doi.org/10.1145/2998181.2998322

[41] Thomas W. Malone and Kevin Crowston. 1990. What is Coordination Theory and How Can It Help Design Cooperative Work Systems?. In *Proceedings of the 1990 ACM Conference on Computer-supported Cooperative Work (CSCW '90)*. ACM, New York, NY, USA, 357–370. https://doi.org/10.1145/99332.99367

[42] Thomas W. Malone, Thomas W. Malone, and Kevin Crowston. 1994. The Interdisciplinary Study of Coordination. *ACM Comput. Surv.* 26, 1 (March 1994), 87–119. https://doi.org/10.1145/174666.174668

[43] Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J. Watts. 2016. An Experimental Study of Team Size and Performance on a Complex Task. *PLOS ONE* 11, 4 (apr 2016), e0153048. https://doi.org/10.1371/journal.pone.0153048

[44] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, New York, New York, USA, 224–235. https://doi.org/10.1145/2531602.2531663

[45] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. 2016. Turking in a Global Labour Market. *Computer Supported Cooperative Work (CSCW)* 25, 1 (01 Feb 2016), 39–77. https://doi.org/10.1007/s10606-015-9241-6

[46] Winter Mason and Duncan J Watts. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '09*. ACM Press, New York, New York, USA, 77. https://doi.org/10.1145/1600150.1600175

[47] J.E. McGrath. 1984. *Groups: Interaction and Performance*. Prentice-Hall. https://books.google.ca/books?id=4pzZAAAAMAAJ

[48] David M. Messick and Keith P. Sentis. 1979. Fairness and preference. *Journal of Experimental Social Psychology* 15, 4 (1979), 418–434. https://doi.org/10.1016/0022-1031(79)90047-7

[49] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359–1373. https://doi.org/10.1287/mnsc.1050.0379

[50] Meredith Ringel Morris, Jeffrey P. Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. 2017. Subcontracting Microwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, New York, New York, USA, 1867–1876. https://doi.org/10.1145/3025453.3025687

[51] Richard T. Mowday. 1979. Equity Theory Predictions of Behavior in Organizations. In *Motivation and Work Behavior* (2 ed.), Richard M. Steers and Lyman W. Porter (Eds.). Mcgraw-Hill, New York, 111–131.

[52] Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica* 1, 1-4 (1986), 251–266. https://doi.org/10.1007/BF01840446

[53] Gary M. Olson and Judith S. Olson. 2000. Distance Matters. *Hum.-Comput. Interact.* 15, 2 (Sept. 2000), 139–178. https://doi.org/10.1207/S15327051HCI1523_4

[54] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *ICWSM International Conference on Web and Social Media* (2011), 321–328. https://doi.org/10.13140/RG.2.2.19170.94401

[55] Michael Ross and Fiore Sicoly. 1979. Egocentric biases in availability and attribution. , 322–336 pages. https://doi.org/10.1037/0022-3514.37.3.322

[56] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2017. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 1700–1713. https://doi.org/10.1145/2998181.2998300

[57] Shikhar Sarin and Vijay Mahajan. 2001. The Effect of Reward Structures on the Performance of Cross-Functional Product Development Teams. *Journal of Marketing* 65, 2 (2001), 35–53. https://doi.org/10.1509/jmkg.65.2.35.18252

arXiv:https://doi.org/10.1509/jmkg.65.2.35.18252

[58] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–19. https://doi.org/10.1145/3274423

[59] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, New York, New York, USA, 937–945. https://doi.org/10.1145/2675133.2675239

[60] Maximilian Speicher and Michael Nebeling. 2018. GestureWiz: A Human-Powered Gesture Design Environment for User Interface Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–11. https://doi.org/10.1145/3173574.3173681

[61] Leigh Thompson and George Loewenstein. 1992. Egocentric perceptions of fairness and interpersonal conflict. *Organizational Behavior & Human Decision Processes* 51 (1992), 176–197.

[62] Ming Yin, Yiling Chen, and Yu-An Sun. 2013. The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[63] Ming Yin, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. The Communication Network Within the Crowd. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. ACM Press, New York, New York, USA, 1293–1303. https://doi.org/10.1145/2872427.2883036

[64] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, New York, New York, USA, 217. https://doi.org/10.1145/2207676.2207708

[65] Sharon Zhou, Melissa Valentine, and Michael S Bernstein. 2018. In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3173574.3173682