CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing

SHAOYANG FAN, The University of Queensland, Australia
UJWAL GADIRAJU, Delft University of Technology, Netherlands
ALESSANDRO CHECCO, University of Sheffield, United Kingdom
GIANLUCA DEMARTINI, The University of Queensland, Australia

Paid micro-task crowdsourcing has gained in popularity partly due to the increasing need for large-scale manually labelled datasets which are often used to train and evaluate Artificial Intelligence systems. Modern paid crowdsourcing platforms use a piecework approach to rewards, meaning that workers are paid for each task they complete, given that their work quality is considered sufficient by the requester or the platform. Such an approach creates risks for workers; their work may be rejected without being rewarded, and they may be working on poorly rewarded tasks, in light of the disproportionate time required to complete them. As a result, recent research has shown that crowd workers may tend to choose specific, simple, and familiar tasks and avoid new requesters to manage these risks.

In this paper, we propose a novel crowdsourcing reward mechanism that allows workers to share these risks and achieve a standardized hourly wage equal for all participating workers. Reward-focused workers can thereby take up challenging and complex HITs without bearing the financial risk of not being rewarded for completed work. We experimentally compare different crowd reward schemes and observe their impact on worker performance and satisfaction. Our results show that 1) workers clearly perceive the benefits of the proposed reward scheme, 2) work effectiveness and efficiency are not impacted as compared to those of the piecework scheme, and 3) the presence of slow workers is limited and does not disrupt the proposed cooperation-based approaches.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**.

Additional Key Words and Phrases: Crowdsourcing, Human Computation, Fairness, Worker behavior, Reward sharing

ACM Reference Format:
Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. 1, 1 (October 2020), 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Micro-task crowdsourcing as a way to acquire large-scale human input has become increasingly popular in research, practice, and various businesses. Requesters deploy Human Intelligence Tasks (HITs) on crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk) [1] and allocate monetary rewards to completed HITs. On the other side of the platform, hundreds of workers

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

Authors' addresses: Shaoyang Fan, fsysean@gmail.com, The University of Queensland, St Lucia, Brisbane, QLD, Australia; Ujwal Gadiraju, u.k.gadiraju@tudelft.nl, Delft University of Technology, Delft, Netherlands; Alessandro Checco, a.checco@sheffield.ac.uk, University of Sheffield, Sheffield, United Kingdom; Gianluca Demartini, g.demartini@uq.edu.au, The University of Queensland, St Lucia, Brisbane, QLD, Australia.

XXXX-XXXX/2020/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

self-select tasks that they wish to complete and receive the associated rewards, if requesters accept their submissions.

Some researchers use micro-task crowdsourcing as a cheap way to obtain high-quality data quickly [4]. Micro-task crowdsourcing has been used to deploy HITs for a variety of purposes such as content moderation and metadata annotations [23]. Other studies have employed micro-task crowdsourcing to design hybrid human-machine methods for improving the quality of machine learning algorithms [13]. Researchers in the HCI, natural language processing and computer vision areas have also generally accepted online micro-crowdsourcing platforms to empower innovative technologies [14, 37, 50, 55]. However, result quality and worker satisfaction are still open research problems [32].

The development of micro-task crowdsourcing is not only crucial for researchers but also plays a significant role for workers on paid crowdsourcing platforms. According to a survey of crowd workers on the MTurk and Figure Eight platforms [2], almost 40% of workers reported crowdsourcing as being their primary source of income, but that this income is usually barely enough to pay for their daily expenses. Furthermore, Hara et al. [31] showed that the median hourly wage for workers on MTurk is only $2 and only 4% of workers achieve more than the US federal minimum wage of $7.25. Improving crowd workers' income and worker satisfaction are essential research challenges as the monetary reward attached to tasks is the primary motivation for people to complete crowdsourcing tasks [3]. Low average hourly wages exist on crowdsourcing platforms as there is a very high number of low-reward tasks and workers willing to complete them [41]. Unpaid work also plays a vital role in decreasing the average wage since working on rejected or abandoned tasks leads to unrewarded work time [28]. More importantly, workers in the crowdsourcing market face risks such as unfair rejection and unexpected losses [42, 45]. These risks not only prevent workers from receiving rewards but also cause the reputation of workers to decrease [21]. The decrease in the reputation caused by such risks directly affects the opportunity for these workers to access more paid work in the future.

In this paper, we propose a novel crowdsourcing reward framework that alleviates worker risks such as unfair and unexpected losses and redistributes rewards across a group of cooperating workers. In order to observe the impact of different reward sharing models, we analyze worker performance and behaviour across five distinct experimental conditions in a between-subjects study. Experiments are carried out on MTurk, which is one of the most popular paid micro-crowdsourcing platforms. https://www.overleaf.com/project/5ed605ac87e8e600015407b7 The conditions we study vary based on whether or not the group of workers are exposed to the proposed reward mechanism, whether or not the monetary incentive is visible to workers in the group, and whether or not workers can observe the performance of other workers in the group. The novel reward distribution models we propose and experimentally evaluate in this paper can provide hourly wages rather than piecework rewards to workers and allow to share the risks of being rejected or poorly paid by requesters, akin to a mutual insurance company. By centrally collecting rewards and distributing them proportionally to the time spent completing HITs, it may even be possible to keep a share of the collected rewards in order to provide crowd workers with sick and holiday leave as well as other social security services like retirement funds.

Our results confirm that under our novel risk-sharing reward scheme, workers benefit from several positive effects (e.g., the ability to focus on completing HIT accurately rather than being as quick as possible to increase their earnings) and requesters obtain result quality levels similar to traditional piecework reward mechanisms. Additionally, we did not observe the presence of free-riders; the influence of slow workers on others' rewards is also limited. In a post-study survey, workers reported a definite interest in participating in cooperative teams with shared risks and rewards, even with slow workers, further validating the positive perception of our proposed reward distribution model.

2 RELATED WORK

The design of platforms like MTurk is unbalanced and favours requesters, who can reject finished tasks without rewarding workers, but can still hold onto the data generated by workers [42]. Preferentially selecting fair requesters to boost earnings and avoiding bad reputation requesters to avoid risks is becoming progressively crucial for crowd workers. However, poor search interfaces and inadequate support for workers create a worse work environment [20, 29]. As a result, workers can only seek mutual help on external websites or tools that can help workers make optimal decisions [40]. In this section we discuss previous work that aimed at supporting crowd worker collaboration and at increasing worker-requester trust.

Workers utilize some tools to make decisions about the tasks to select to work on and deal with the existing information asymmetries that exist in crowdsourcing marketplaces. In 2008, Silberman and Irani designed and launched Turkopticon: a platform that is employed by MTurk workers to review requesters. Thanks to a browser extension that aggregates requesters' review data and displays the aggregated metrics directly on MTurk, such tool exposes workers to reputation scores for a requester before accepting to complete their HITs [48]. One limitation of this approach is that requesters who make mistakes and receive bad review scores may not be able to access quality workers in the future. Also, there is no method to control for false reviews.

Callison-Burch developed the Crowd-workers browser plug-in to automatically collect mutual aid data as workers complete HITs, such as the time it takes for workers to complete tasks and the reward workers receive [5]. The system then aggregates the collected data and calculates the HIT average hourly rate, which allows workers to find high-paying tasks. The results depend on past data and involving just a few workers would lead to inaccurate measurements. Besides, the lack of an estimated hourly wage can lead to new requesters with no previous data being perceived as less attractive by workers.

In addition to helping workers make HIT choices, some workers self-organize into independent online communities. For instance, the online forum TurkerNation provides workers with tips and strategies to work efficiently on MTurk. Ma et al. explored the community side of crowd workers in online communities and showed that the active participation of workers in these forums reduces the desire of leaving the crowdsourcing platform [39].

In summary, existing tools and platforms help workers in mitigating risks by avoiding bad tasks and requesters in the short term. However, these tools have some drawbacks. The first drawback is that these mutual aid tools usually have a particular hostility towards the requesters, which may make the worker-requester trust relationship worse [42]. The second drawback is their vulnerability. For example, some bad requesters may re-create new accounts to cover up bad reputations while some new potentially good requesters may not be able to attract good workers because of early HIT design mistakes one can associate with the general learning curve.

In the long term, such tools may harm the future development of crowdsourcing platforms. Therefore, a mechanism that can alleviate risks and build trust between requesters and workers is critical. To support worker-requester trust development, Salehi et al. developed Dynamo, a platform to support the MTurk community with collective actions. As an example, a campaign that is called "Guidelines for Academic requesters" in Dynamo successfully involved both workers and researchers supporting fair treatment of crowd work. However, Dynamo faces barriers for collective action because of stalling and friction [47]. McInnis et al. provided several practical task design suggestions to support effective communication between workers and requesters and to mitigate the risk of unfair rejection [42]. Chang et al. applied some of the design suggestions to decrease unfair rejections by allowing workers to present their diversity and creativity in the process and to offer feedback in later stages [6]. However, this system is limited to data labelling tasks.

Along this line of research, in this paper, we develop a different mechanism to foster mutual trust. First of all, substantial research focused on unfair rejection, studying how unfair rejection negatively impacts workers' performance [21, 26, 42]. However, unfair rejection is only a part of the risks that workers face in the micro-task crowdsourcing market. We study the nature of risks from quality control mechanisms, unfair rejection and traditional piece work reward schemes.

We allocate workers into a group that share risks, including rejection and poorly paid HITs, among group members. Although workers can avoid bad requesters and HITs through external websites or tools, these methods build a barrier for new requesters. Such optimizations can also lead to workers paying less attention to the quality of their work and more on how to make money with less effort in a short time [7]. Our alternative reward mechanism allows workers to concentrate their efforts on completing tasks *effectively* rather than *efficiently*. Co-op reward schemes remove the pressure on workers to work efficiently and complete the task as quickly as possible to maximize their reward – behavior that has typically been exhibited corresponding to piecework reward schemes [22]. Our method also encourages workers to try new and different high-risk tasks. Even reward-motivated workers can take up impressive, possibly challenging and complex HITs without worrying about the associated risks.

3 WORKERS' RISKS AND REWARD DISTRIBUTION MODELS

In this section, we first discuss the risks that workers frequently encounter in the micro-task crowdsourcing market. We then introduce novel reward distribution models which we then experimentally evaluate together with crowd workers in Section 5.

3.1 Risks Generated by Requesters

Requesters may reject workers' job due to bad quality, but some work may be rejected unfairly [30, 42, 45]. Prior work showed that some requesters reject work deliberately without a justifiable reason [42]. Moreover, technical errors made by requesters may cause task submission to fail thus leading to abandonment and unrewarded work time. Requesters on a micro-task crowdsourcing platform need to design their tasks. However, some requesters lack experience, so tasks may not be well designed or may have confusing instructions [25]. Such poor design increases the difficulty of the task, sometimes causing workers to give up halfway [28]. These risks not only prevent workers from receiving rewards but also cause workers' reputation to decrease. This directly affects the opportunity for these workers to access more tasks in the future.

Requesters improve the quality of crowd work by employing external quality control methods such as pre-screening questions or gold questions (i.e., tasks with verifiable answers) to monitor and filter out malicious workers [12, 49]. The use of gold questions, if designed poorly, can increase the chance of mistakes and can have detrimental effects on workers. For instance, some instructional attention check questions are complicated to understand for workers and result in unintentional mistakes [24]. We observed also in our experiment that workers sometimes missed attention check question, but the remainder of their work in the batch was still valuable and of good quality. This unpaid work not only decreases the average wage but also hurts workers' motivation. A recent study has also shown that workers who receive rejection exhibit negative emotional reactions [21].

3.2 Risks Generated by Platforms

Requesters on a crowdsourcing platform typically set the payment for each HIT. They usually follow two methods to determine the HIT reward. The first is based on market pricing, adapting the reward to similar HITs running on the same platform. For example, information finding tasks on MTurk are often paid \$0.02 or \$0.03 per HIT, but sometimes the task could take up to 5 minutes to complete [13]. A different approach to set rewards is based on the time needed to complete the HIT through rapid prototyping [8]. Requesters complete HITs or publish a sample of HITs on MTurk to estimate how long it takes to complete. Next, the reward for each HIT is calculated based on a set hourly wage (e.g., the requester's country minimum wage). Once the requester has set the individual HIT reward, the total reward that workers receive is computed considering the sum of the rewards allocated to all the HITs the worker has completed on the platform. Recently, by adding a line of script to task HTML on MTurk, Whiting et al. [52] introduced Fair Work to enable workers to self report task duration and then automatically pay minimum wage to workers. Even if requesters set HIT rewards based on a minimum hourly wage, the traditional piecework reward model comes with risks. Questions have different inherent difficulty levels, so various tasks in the same batch may take different amounts of time to complete. For example, workers who need to annotate images of machinery components with model information may face varying levels of task difficulty based on the camera angle and clarity of each image. Some may be very blurred, leading to workers being unable to provide sufficient or accurate information. This situation goes against the fairness of piecework reward schemes. A higher level of task difficulty leads to lower worker accuracy and longer task completion time. In the case of the traditional reward distribution methods (called *Baseline* in our experiments), HITs of different difficulties that are bundled together in a batch can have an adverse effect on worker motivation. This is due to the varying amount of time required for task completion despite the associated reward being the same across all HITs.

Most current crowdsourcing platforms collect workers' historical performance metrics to measure the workers' reliability. Requesters can pre-select workers by setting a threshold on the historical acceptance rate. In order to keep a good track record, workers start to pick easy and familiar tasks. This creates a new risk for workers as they lose the chance to learn a new skill. All of these cases increase the risk for workers to be underpaid or not paid at all in case of rejection. Those risks are inevitable as the platforms create them. However, the novel reward distribution models that we introduce in this paper aims at sharing those risks across team members of a group.

3.3 Reward Distribution Model

We propose a novel crowdsourcing reward mechanism that adds participating workers into a co-op group to share risks. The reward that workers in the group receive is computed considering the sum of the rewards allocated to the HITs workers in the group had completed. According to *co-op payments*, individual workers are paid based on the *amount of time* that they spent completing HITs. The entire co-op group earnings are thus redistributed among its members in proportion to the time they spent on HITs. This allows workers to share the risks of working on poorly paid HITs and makes the hourly wage the same for all members of the co-op group.

Comparing the piecework and co-op reward schemes, we can observe how the individual optimization approach aiming at completing HITs quickly to increase the personal hourly wage disappears in the proposed co-op reward scheme where workers can focus on producing accurate labels for requesters without worrying about single HIT reward. At the same time, missed reward due to rejection can be avoided by sharing the loss among group members, thereby diluting the impact originally put on individual workers whose work has been rejected. Efficient crowd workers who contribute most reward to the group will cover for less experienced workers who may be initially slower or less accurate in completing HITs.

Table 1. Different reward distribution conditions.

	Shared Reward	**Visible Incentive**	**Social Comparison**
Baseline	✗	✓	✗
HidIn	✗	✗	✗
CooVi	✓	✓	✗
CooHi	✓	✗	✗
CooHiSo	✓	✗	✓

4 METHOD AND EXPERIMENTAL SETUP

4.1 Study Design

In order to observe the impact of different reward sharing models, we collect and analyze worker behavior and performance data across five different reward distribution conditions. Table 1 presents the variations across the different conditions.

Piecework with Visible Incentive (Baseline). Workers are asked to complete HITs on an external platform we developed to be similar to MTurk where the reward and the number of available HITs are displayed on the task list page. Under this condition, workers are paid on a piecework basis like on MTurk and are free to pick which HIT batch from the list of available HITs they want to work on. The total reward that they will receive is updated after each completed HIT and is displayed on the page. This condition mirrors the traditional piecework reward model, and there is no risk and reward sharing among workers.

Piecework with Hidden Incentive (HidIn). Under this condition, the reward and batch size for HITs are hidden in the task list page. Therefore, workers select HITs to complete based on their own interest rather than on the reward attached to them. Workers are still rewarded on a piecework basis.

Co-operative with Visible Incentive (CooVi). Under this condition, the reward that workers receive is based on the amount of time they spent completing HITs as explained in Section 3.3. The entire co-op group earnings are redistributed among its members in proportion to the time spent. The reward allocated to individual HITs and the HIT batch size are visible to workers.

Co-operative with Hidden Incentive (CooHi). Under this condition, the reward-sharing scheme is the same as in the *CooVi* condition, except that the individual HIT reward and the HIT batch size are not visible in the task list page.

The combination of *Baseline*, *HidIn*,*CooVi* and *CooHi* consists of a 2 x 2 factorial design to test the effect of reward-sharing and the effect of visible reward information on workers' performance. It also examines the interaction effect of these variables. For example, it determines whether the reward-sharing scheme influence workers' performance when they choose tasks according to their interests instead of the reward level. Overall, there are three hypotheses we aim to test with such design.

Null Hypothesis 1. *Workers' performance is not affected by using a reward-sharing scheme.*

Null Hypothesis 2. *Workers' performance is not affected by reward transparency.*

Null Hypothesis 3. *There is no interaction effect between reward-sharing and visible incentives.*

Co-operative with Hidden Incentive and Social Comparison (CooHiSo). We add one experimental condition to additionally investigate how the visibility of other group members' performance

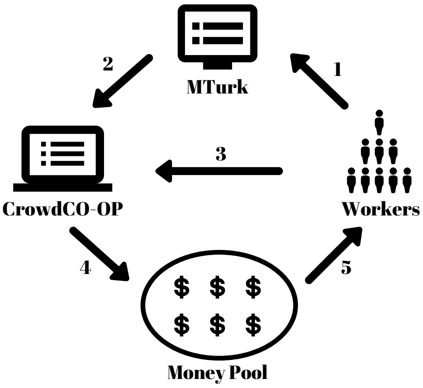


Fig. 1. Overview of CrowdCO-OP: (1) workers are recruited from MTurk; (2) Workers are redirect to CrowdCO-OP; (3) Workers select and finish tasks on CrowdCO-OP; (4) the rewards allocated to completed HITs are collected into a money pool; (5) the rewards are redistributed to worker based on work time.

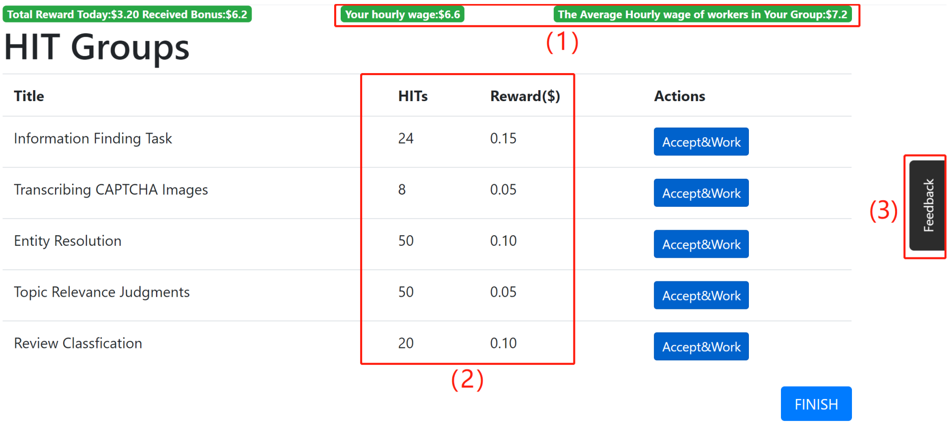
affects worker behavior in a co-operative setting without visible rewards. Festinger introduced the theory of social comparison in 1954, which describes people's self-evaluation done in the absence of objective means (hidden incentive, in our case) by comparing their attitudes, abilities and beliefs with others [19]. The reward-sharing scheme and invisible incentive condition are the same as in *CooHi*, except that the hourly wage of other workers in the group is also displayed on the page to serve as means for social comparison. The results from *CooHi* and *CooHiSo* can be used to test the following hypothesis.

Null Hypothesis 4. *The existence of social comparison does not affect workers' performance when HIT information is not transparent in a co-operative setting.*

4.2 Experimental Setup

We leveraged the MTurk API in a similar way to TurkPrime [38] to build an online crowdsourcing platform called **CrowdCO-OP**. Figure 1 shows a sketch of our method. CrowdCO-OP is similar to MTurk and TurkPrime but for the type of reward distribution and reward information presented to workers over the user interface which can be controlled across different conditions. Unlike TurkPrime, we only recruit participants and send rewards through the MTurk platform in the form of bonuses. Other functions, such as task design and worker responses, are completely independent from the MTurk platform. The corresponding data (i.e., behavioural logs and task answers) is entirely stored in a separate database on a dedicated server.

We published five HIT batches on MTurk representing the five different reward distribution conditions presented in study design, and recruited 50 distinct workers in each condition. Participating workers were limited to the US population since Difallah et al. showed how in recent years, US-only HITs occupied the largest share of the MTurk market in terms of quantity and incentives [13]. To prevent biases due to learning effects, workers were not allowed to participate in more than one condition. Additionally, the IP addresses of workers were logged to prevent workers using multiple accounts. In each HIT, workers were first asked to complete a demographics questionnaire with questions about their gender, age, education and political views. After submitting the questionnaire, workers were redirected to CrowdCO-OP where they could select HITs to complete from the task list page. On completing the survey, workers received a completion code that they could use to



Total Reward Today:$3.20 Received Bonus:$6.2

Your hourly wage:$6.6

The Average Hourly wage of workers in Your Group:$7.2

HIT Groups

Title	HITs	Reward($)	Actions
Information Finding Task	24	0.15	Accept&Work
Transcribing CAPTCHA Images	8	0.05	Accept&Work
Entity Resolution	50	0.10	Accept&Work
Topic Relevance Judgments	50	0.05	Accept&Work
Review Classfication	20	0.10	Accept&Work

Fig. 2. Screenshot of the task list page on CrowdCO-OP: (1) information about other group members' performance; (2) reward and batch size for HITs; (3) feedback form.

submit the HIT on MTurk. The submitted task was automatically approved when a valid code was entered.

On CrowdCO-OP, workers in all conditions are first presented with a page that is similar to the available task list page on MTurk. Figure 2 shows the task list page on CrowdCO-OP. For each HIT batch, they can see a title and a description of the task, the reward attached to the HIT, and the batch size. Metrics about the performance of other workers in the group were only available in *CooHiSo*. Workers could freely interact with the HITs, preview them, and decide which ones to work on. Workers were allocated a maximum of six minutes to complete each of the tasks and there was no penalty for abandoning.

The rewards earned by workers on our external CrowdCO-OP platform are sent to them as a bonus on MTurk within 24 hours from the completion of their work session. Workers were also informed about the possibility of returning to the list of tasks available to them anytime during the following days (up to a maximum of 20 days) by using their personal URL which identifies them on CrowdCO-OP.

4.3 Onboarding Workers

While involving MTurk workers in co-op groups, a key aspect is explaining them how they will be rewarded for their work towards completing HITs. This allows them to be aware of the implications of risk and reward sharing and impacts the way in which they approach HIT completion. While, on the one hand, workers aware of participating to a co-op group can take the necessary time to complete HITs accurately, on the other hand, some may leverage the cooperative reward scheme setup to be paid without being productive (i.e., *free-riders*). Note, however, that this is not different from workplaces that require employees to clock in and out and are paid accordingly to the hours they worked. Such employees when being unproductive face work appraisal processes which may lead to their contract being ended. While we look at the presence and impact of free-riding in this paper, we envision a similar approach being possible to deal with free-riding in co-op groups.

In our experiment, we recruited workers from MTurk, presented them with an explanation including examples customized for the different condition they were assigned to, and paid them the accrued reward as a bonus on MTurk. Before starting the experiment, workers have provided informed consent to the study. The five recruitment HITs (one per condition) have been posted on the MTurk platform, and workers could choose whether to participate in the experiment or not after being explained the experimental conditions. However, when a worker agreed to join one experimental condition, they were not be able to accept the HIT for the other four experimental conditions. In the end, the total reward that workers received was $0.01 allocated to the recruitment tasks posted on MTurk plus the reward based on the tasks they did on CrowdCO-OP sent to them via a bonus on MTurk.

4.4 Tasks Design

To simulate the diverse types of HITs available on the MTurk platform, we deployed several task types including Content Creation (CC), Information Finding (IF), Interpretation and Analysis (IA) and Verification and Validation (VV) tasks [13]. Each batch contained tasks with varying complexity, wherein some tasks required relatively more time to complete and workers had a higher risk of committing errors. However, replicating MTurk, every task within a batch was associated with the same monetary reward. The batches of HITs on the task list page, as well as individual tasks within the batch were randomized to avoid bias due to ordering effects. The HIT batches workers were presented with on CrowdCO-OP are described below.

Table 2. Different batches of HITs that workers were presented with across all conditions.

Task Type	Units per HIT	HITs per Batch	HIT Reward
Information Finding [22]	5	24	$0.15
CAPTCHAs [22]	5	8	$0.05
Entity Resolution [51]	5	50	$0.1
Relevance Judgment [16]	1	50	$0.05
Review Classification [9]	5	20	$0.1

Information Finding Tasks (IF). Workers were asked to find the middle name of 5 given people. These tasks had three levels of difficulty, and with the increasing level of difficulty, workers needed to consider an additional constraint in the information finding process. In the simplest task, the workers only needed to find the middle name of a unique celebrity. At the second level of difficulty, workers needed to consider the profession of the person to disambiguate between several individuals with the same name. In the third level of difficulty, workers needed to deal with two steps of disambiguation to determine the middle name of the person they were looking for; according to the year in which the person was active in the given profession.

Transcribing CAPTCHA Images (TCI). Workers were asked to decipher characters from five given images. These tasks had a different level of difficulty depending on the number of strokes that were used to smudge the CAPTCHA. Less interfered images could be transcribed more easily than other images.

Entity Resolution (ER). For each pair of products, workers needed to verify whether they referred to the same product or not. The information of the product was integrated from two different sources and each record had two attributes: name and price.

Topic Relevance Judgment (TRJ). Workers needed to read one document to determine whether it was relevant to the given topic. The topic was, "*birth rates are falling in other countries besides the United States and China*". For example, a document was relevant to the topic because it described that Japan had lower birth rates than in the past.

Review Classification (RC). Workers were presented with five reviews related to fashion items and they needed to classify reviews into one of the following three classes: size issue, fit issue, and no issue with size or fit. The first category pertained to negative feedback on the size, in comparison to the normal size. The second category was about comfort. The last category corresponded to feedback which had nothing to do with either size or fit.

5 RESULTS

The recruiting tasks for each group were published on the MTurk platform between March 2019 and April 2019. 250 workers submitted valid completion codes and completed at least one HIT on CrowdCO-OP. As described earlier, all workers had 20 days to carry out a maximum of 152 HITs on CrowdCO-OP after accepting the task on MTurk. In total 13'084 HITs were completed by workers on CrowdCO-OP, while 51'734 answers and 78'470 log records of worker activities were collected from 250 workers. On average, every worker completed 52.33 HITs (*SD=54.70; median=25*) and spent 3'682.84 seconds (*SD=4571.51; median=1707.00*) on our platform.

Table 3. Worker demographics for each condition.

	Baseline	*HidIn*	CooHi	*CooHiSo*	*CooVi*
Median Age	33	32	33	37	34
% Female	52%	70%	76%	52%	54%
% Went College	70%	78%	72%	70%	64%
% Married	40%	28%	42%	40%	40%
% Democrat	52%	44%	34%	30%	32%
% Income Under $50,000	54%	56%	48%	50%	54%
% Political View as Moderate	30%	44%	34%	40%	30%
% MTurk Exp Under 6 Months	36%	50%	44%	40%	64%

Table 3 summarizes the demographic information reported by workers. Workers across the different conditions share some demographic characteristics. Workers across all conditions were mainly between 30 and 40 years old, well-educated and earning less than $50,000 a year. We note small gender differences in the *Baseline*, *CooHiSo* and *CooVi* conditions, but *HidIn* and *CooHi* are dominated by more than 70% female workers. The proportion of married workers in all conditions except *HidIn* is also similar at about 40%. Above 30% of the workers identify their political view as `moderate` and over half of the workers in the *Baseline* condition self-report as `Democrats`. *CooVi* and *HidIn* have a large number of novice workers; about 64% of workers in CooVi and half of the workers in HidIn have less than six months experience on the MTurk platform. Interestingly, we also note that the majority of workers in co-op reward groups were liberal despite the onboarding activity (where we explain how the reward mechanism works) was done before the demographics questionnaire and workers could still decide not to participate.

As the demographic diversity in each group may affect the experimental results, we ran Spearman's rank-order correlation tests to assess the correlation between worker demographics and performance metrics such as the number of completed HITs, the accuracy rate, and the average HIT completion time. All Spearman correlation coefficients shown to be very low showing that

Table 4. Overall worker performance across all groups.

	Baseline	*HidIn*	*CooHi*	*CooHiSo*	*CooVi*
# **HITs**	3688	1844	1519	2577	3456
# **Tasks**	14851	7140	6316	9926	13501
Accuracy	82.83%	82.93%	85.12%	81.96%	86.77%

correlations between demographic variables and workers' performance are negligible. Therefore, we can ignore the association between worker demographics and experiment results.

5.1 Do Different Reward Allocation and Visibility Schemes Influence Workers' Performance?

Table 4 shows the overall worker performance under each condition. It is clear that the number of HITs completed by the workers in the *Baseline* and the *CooVi* groups were significantly higher than others (3'688 and 3'456 respectively), whereas workers in the *HidIn* completed the least number of tasks. We investigated Null Hypotheses 1, 2 and 3 based on the number of HITs completed by workers. To this end, an aligned ranks transformation ANOVA (ART ANOVA) test [54] was conducted due to the non-normally distributed data and the existence of outliers. The interaction effect between Shared Reward and Visible Incentive in the number of HITs was not statistically significant, $F(1, 195) = 0.342, p = .559$, partial $\eta^2 = .002$. Therefore, an analysis of the main effect for Visible Incentive was performed, which indicated that the main effect was statistically significant, $F(1, 195) = 34.35, p < .001$, partial $\eta^2 = .15$. We run a post-hoc pairwise comparison with Tukey adjustment. The marginal means for the number of HITs were 78.1 ± 5.3 for the Hidden Incentive and 122.1 ± 5.33 for the Visible Incentive, a statistically significant mean difference of 44.1, $p < .001$. This post-hoc analysis shows that workers who were shown the associated HIT reward and the number of available HITs (in the *Baseline* and *CooVi*) completed more HITs than workers without visible incentives (in the *CooHi* and *HidIn*). Furthermore, the main effect of reward-sharing in the number of HITs was not statistically significant, $F(1, 195) = 0.007, p = .933$, partial $\eta^2 = .00003$. So, independently on whether information about HITs is presented to workers or not, the number of HITs completed by workers in groups using the co-op reward distribution method was not significantly different.

We also looked at worker accuracy rates in relation to our main hypotheses. Table 4 shows that workers in *CooVi* exhibit the highest average accuracy rate of 86.74% while workers in *CooHiSo* have the lowest average accuracy rate of 81.91%. An ART ANOVA test was carried out to test null hypotheses 1, 2 and 3 based on accuracy due to the non-normally distributed data. All three null hypotheses were accepted due to a lack of statistical significance. for shared reward, $F(1, 195) = 0.85$, $p = 0.358$; for visible incentive, $F(1, 195) = 3.559$, $p = 0.06$; and for interaction effects, $F(1, 195) = 2.036$, $p = 0.155$. Therefore, although CooVi showed to result into better overall accuracy, there was no statistically significant difference in accuracy across interventions. This shows how the different reward schemes had no impact on the quality of labels generated by the crowd. Specifically, co-op reward sharing mechanisms do not decrease the quality of crowd work despite the benefits they bring to workers in terms of risk management.

By comparing quantity and quality in the *CooHi* and *CooHiSo* conditions using Mann-Whitney U tests, we found that all differences were not significantly significant ($U = 1518$, $Z = -1.849$, $p = 0.064$ and $U = 1314.5$, $Z = 0.658$, $p = 0.51$, respectively). Therefore, Null Hypothesis 4 is not rejected. Under reward-sharing conditions, the use of social comparison does not affect the workers' performance when HIT information is not transparent.

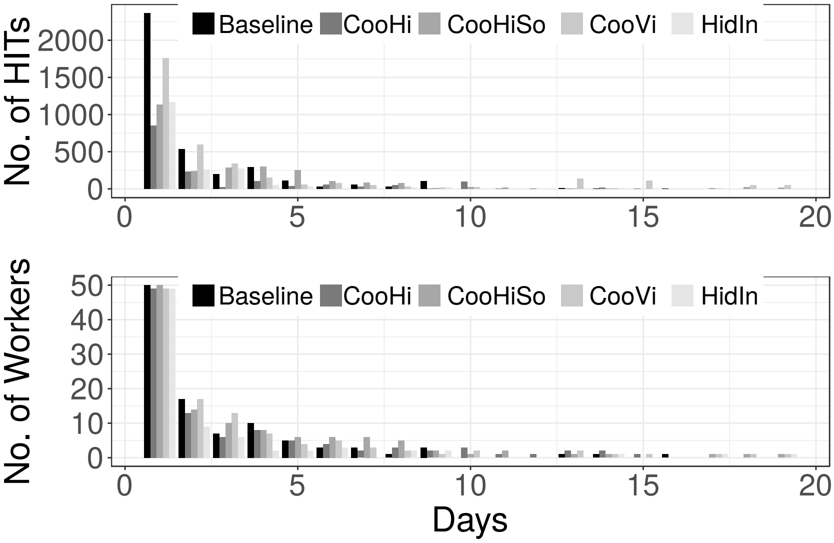


Fig. 3. Number of submitted HITs and workers across different conditions over 20 days.

Key Findings. First, workers completed more tasks when they were presented with more information about HITs, regardless of whether they shared their rewards or not. This means that visible rewards and platform transparency play a significant role in encouraging workers to finish more tasks. The different reward and risk sharing schemes did not have a significant impact on the number of tasks completed by workers. This suggests that the proposed reward distribution methods can create a win-win situation; workers' risks are reduced, and their enthusiasm to complete tasks is not affected (thus benefiting requesters).

5.2 Do Different Reward Distribution and Visibility Schemes Influence Worker Performance Over Time?

We analyzed how workers approached the HITs across the 5 conditions. Figure 3 illustrates the number of finished HITs and number of workers across the different conditions over 20 days. Dates were normalized according to the initial task acceptance time of each participating worker on MTurk. For instance, after a worker accepted the HIT that was released through the MTurk platform, the tasks completed on our platform over the next 24 hours were counted as the HITs completed on their first day. Interestingly, we found that most workers completed a large number of HITs on the CrowdCO-OP platform on their first day.

In this section, the data of dependent variables for each group of independent variables are not normally distributed, and the interaction effect is not considered. So we ran all combinations of factor levels as one factor non-parametric test with five levels to facilitate making the group comparisons over days. Table 5 provides p values of Kruskal-Wallis H Tests in comparing worker performance and average HIT duration (AHD) across five groups over 20 days. We merged data from the 6th day to the 20th day because active workers reduced dramatically after the 5th day. From Table 5, we can observe statistically significant differences in the number of HITs completed on their first day ($x^2(4) = 24.257$, $p < .001$), third day ($x^2(4) = 13.144$, $p = 0.011$) and during the last few days ($x^2(4) = 11.265$, $p = 0.024$) across five groups. On the other hand, the null hypothesis is not rejected for the accuracy and AHD.

Table 5. p-values of Kruskal-Wallis H Tests in worker performance and average HIT duration (AHD) across all groups over 20 days

	Day 1	*Day 2*	*Day 3*	*Day 4*	*Day 5*	*Day 6-20*
# **HITs**	< 0.001	0.144	0.011	0.186	0.382	0.024
Accuracy	0.556	0.462	0.213	0.101	0.304	0.442
AHD	0.174	0.27	0.065	0.995	0.979	0.121

Null hypothesis: There were no differences in worker performance between groups that differed in reward distribution and visibility schemes

Some interesting observations can be made based on the post-hoc analysis with a Bonferroni correction for multiple comparisons. First, statistical differences in the number of HITS completed on the first day across conditions followed the same pattern as the overall performance. This post hoc analysis exposed statistically significant differences in the number of completed HITs between *CooHi* (*mean rank=101.10*) and *CooVi* (*mean rank=143.13*); $p = .036$, *CooHi* and *Baseline* (*mean rank=157.61*); $p = .001$, *HidIn* (*mean rank=105.79*) and *Baseline*; $p = .003$, but not between any other condition combination. Workers in the *Baseline* (2'365 HITs) and *CooVi* (1'761 HITs) conditions completed a higher number of tasks on the first day than workers in *HidIn*, *CooHi* and *CooHiSo* (1'166 HITs, 848 HITs and 1'135 HITs respectively). This suggests that the visible incentives associated with HITs in the *Baseline* and *CooVi* conditions spurred workers to complete more tasks on the first day. This is likely due to the familiarity of workers with the piecework model and with visible incentives.

However, this pattern was not observed in the following days and differences in the number of completed HITs only existed on the third day between *CooHi* (*mean rank=6.50*) and *HidIn* (*mean rank=31.17*), $p = .005$. This happens as workers seem to still prefer the traditional reward model when reward information is not visible. Moreover, the post hoc analysis on the last few days revealed statistically significant differences in the number of completed HITs between *CooHi* (*mean rank=35.61*) and *CooVi* (*mean rank=60.27*), $p = .027$, but not between other group combinations. In the long run, the impact of reward information transparency was weak in the traditional reward distribution model but it still played an important role in the reward-sharing schema.

We found that the co-op groups that shared rewards and risks had more active workers during last 10 days of the experiment as compared to workers in conditions where rewards were not shared. For example, during the last 10 days, 16 active workers in *CooHi* returned to our platform using their personal link to complete more HITs, but only 3 workers did so in *HidIn.* While 21 workers in *CooVi* returned, only 5 workers did so in *Baseline.* This suggests that with the same HIT information being visible, the strategy of sharing rewards and risks among a group of workers motivated more workers to return to HIT batches than in the traditional piecework reward method.

Key Findings. We found that workers were quick to accept our new CrowdCO-OP platform, but were skeptical about the reward distribution models and hidden incentives early in the experiment. In the long run, the impact of information transparency was weak in the traditional reward distribution model but still played an important role in the reward-sharing schema as reward-sharing made income assessment more challenging. We also found that co-op reward schemes better incentivizes workers to return to the platform over time.

5.3 Do Reward-Sharing Workers Put More Effort on Completing Tasks Accurately?

We analyzed the impact of task type on HIT completion across the different conditions. The most popular task type on average is Transcribing CAPTCHA Images (TCI); about six-in-ten (57.7%) of the available TCI HITs were finished by workers in 20 days, compared to 35.48% of Information-Finding (IF) tasks, 31.45% of Entity Resolution (ER) tasks, 27.44% of Topic Relevance Judgment (TRJ) tasks and roughly half (51.3%) of Review Classification (RC) tasks. On the contrary, the least popular task type was TRJ.

Table 6 shows a comparison of the number of workers who chose a particular type of task as their first HIT across the 5 conditions. We found that a majority of workers in CrowdCO-OP chose to attempt IF tasks (107) and TCI tasks (84) first, while entity resolution tasks (8) were the least preferred choice[2]. Workers across the different conditions exhibited similar preferences among the task types, prioritizing IF and TCI tasks.

More workers chose IF as their first task than TCI, but TCI was the most popular. Assuming workers select HITs based on interest, an explanation of this result is that workers can transcript CAPTCHA from images without additional information, but IF tasks require web search skills and abstract thinking. Besides this, we conclude that batch size is not another factor influencing the choice of first task as information about the batch size is not available in *HidIn*, *CooHi* and *CooHiSo* but those group display the same phenomenon.

Table 6. Number of times a task has been chosen first by a worker across the different conditions.

	IF	**TCI**	**ER**	**TRJ**	**RC**
Baseline	20	15	4	3	8
HidIn	23	16	1	6	4
CooHi	22	14	1	4	9
CooHiSo	20	17	1	7	5
CooVi	22	22	1	2	3
Overall	107	84	8	22	29

Table 7 shows accuracy and task completion time across task types and conditions. We observe that the average HIT completion time of CooVi was longer than that of *Baseline* except for TCI tasks while the overall accuracy rate of answers from *CooVi* was higher than the accuracy rate of Baseline. We used Mann-Whitney U tests to check whether those differences were significant. Results are displayed in Table 8. For IF and ER, workers in *CooVi* (mean rank=31.63 and 63.96) have

[2]Note that we randomized the order of HIT batches presented in the initial task list page across workers.

Table 7. Accuracy and average completion time per HIT across the different task types.

	IF		TCI		ER		TRJ		RC	
	Accuracy (%)	Avg. Time Per HIT	Accuracy (%)	Avg. Time Per HIT	Accuracy (%)	Avg. Time Per HIT	Accuracy (%)	Avg. Time Per HIT	Accuracy (%)	Avg. Time Per HIT
Baseline	76.04	127.91	70.77	43.60	80.16	47.56	75.55	35.54	93.89	51.03
HidIn	81.42	157.13	75.15	43.43	83.07	50.40	73.29	42.82	90.93	50.69
CooHi	82.44	150.49	73.41	51.66	84.71	48.31	74.34	41.83	95.00	54.01
CooHiSo	74.00	183.83	70.20	51.76	80.25	82.49	78.03	74.29	96.12	64.71
CooVi	86.56	151.90	72.47	44.33	84.62	56.30	77.28	43.53	96.70	51.68
Average	80.09	154.25	72.40	46.96	82.56	57.01	75.70	47.60	94.53	54.42

Table 8. The Mann-Whitney U tests in accuracy and task completion time between Baseline and CooVi.

	Avg Time Per HIT			Accuracy		
	U	*Z*	*p* value	*U*	*Z*	*p* value
IF	117	-3.526	.000	125.5	-3.352	.001
TCI	32	.000	1.000	27	-.525	.600
ER	577	-4.640	.000	859.5	-2.692	.007
TRJ	970	-1.930	.054	1119	-.904	.366
RC	193	-.189	.850	83.5	-3.155	.002

Null hypothesis: There were no differences in accuracy and task completion time between Baseline and CooVi.

longer average HIT completion time per HIT than workers in *Baseline* (mean rank=17.38 and 37.04), $p < 0.001$ and $p < 0.001$ respectively.

IF and ER tasks are more complicated and require more work. For example, workers need to retrieve information from the internet, and only then they can answer questions. More effort to search the web for information will lead to more accurate answers. As a result, we found that for the IF, ER, and RC task types, workers from *CooVi* (mean rank=31.27, 58.31 and 26.33 respectively) have higher accuracy than worker from *Baseline* (mean rank=17.73, 42.69 and 14.68, respectively), $p = 0.001$, $p = 0.007$, $p = 0.002$ respectively. The performance of *CooVi* group workers in the IF, ER and RC task was better than *Baseline*. When the visible incentive is available, workers in the reward-sharing groups put more effort (longer average HIT completion time) into IF and ER tasks. This can be explained by the fact that the distribution of income is related to the time spent on tasks rather than the number of tasks completed. It is worth noting that *CooHiSo* has the opposite trend. In the task of IF tasks, the longer working time leads to the lower accuracy. This happen due to the existence of cheaters, which will be explained in the next section.

On the other hand, the TCI task is straightforward, and workers can get the answer directly instead of needing extra time to look for information. Spending more time on TCI task will not improve performance. That is why the differences in the average task completion time and performance between *CooVi* and *Baseline* are not significant for this task.

Key Findings. These results support our assumption that co-op reward schemes in the visible incentive setting remove the pressure on workers to work efficiently and to complete the task as quickly as possible to get as much reward as possible, as typically done for piecework reward schemes. Our alternative reward mechanisms allow workers to concentrate their efforts on completing tasks effectively rather than efficiently since the distribution of rewards is a function of the time spent completing tasks.

5.4 Does Reward-Sharing Trigger Unfair Behaviours?

On analyzing the task completion time of workers, we found that some workers took an abnormally short or long time to complete tasks. 1’036 tasks were completed in less than one second, with a corresponding close-to-random accuracy of 53% on average. The extremely short task completion times could be a result of malicious activity. We can see from Table 9 that the *Baseline* group had the largest number of workers with abnormally short task completion time at 24. Most importantly, workers in the *Baseline* group completed 575 tasks with unusual completion time. This number was almost six times the number of tasks completed by *CooVi* group workers (95 tasks). The average accuracy rate for the corresponding tasks in the *Baseline* group (59.3%) was also lower than the average accuracy rate for tasks in the *CooVi* group (87.37%). Furthermore, *HidIn* group (114 tasks)

Table 9. The number of workers with unusually short task completion time (less than one second), the number of corresponding tasks and overall accuracy.

	#Workers	#Tasks	Accuracy
Baseline	24	575	59.30%
HidIn	13	114	83.33%
CooHi	12	24	75%
CooHiSo	13	228	36.4%
CooVi	18	95	87.37%

Table 10. Share of low quality workers for different threshold values for gold question accuracy across the different task types.

	Threshold	IF	TCI	ER	TRJ	RC
Baseline	70%	8.11%	2.50%	5.71%	6.25%	2.78%
CooVi		2.50%	0.00%	0.00%	3.45%	0.00%
Baseline	80%	8.11%	10.00%	5.71%	6.25%	2.78%
CooVi		2.50%	5.00%	0.00%	3.45%	0.00%
Baseline	90%	8.11%	10.00%	11.43%	12.50%	2.78%
CooVi		2.50%	5.00%	0.00%	6.90%	2.63%

had about five times more tasks with less than one second completion time than *CooHi* group (24 tasks) had. Since the rewards in the reward-sharing approach are paid out based on the time spent, malicious workers with a short completion time will receive only a tiny proportion of rewards. This outcome is exactly the opposite of the purpose of the cheater as they aim to obtain substantial financial rewards generated by time efficiency in piecework schemes.

Another noteworthy fact is that there were fewer cheaters in groups with opaque information as obscure information prevents them from evaluating how much money they can make. The *CooHiSo* (228 tasks) group also had a higher number of tasks with cheating behaviours than the *HidIn* group, and the accuracy rate of the *CooHiSo* group was the lowest, reaching 36.4%. The reason for this is that, despite the opacity of the information, cheaters can still see the expected income as the average group income is visible in the *CooHiSo* group. We found that on the first day, the average salary of the *CooHiSo* group was as high as $7.16 per hour, which made it attractive to cheaters.

Like quality control methods used by requesters in all crowdsourcing platforms, we simulated the use of gold questions to identify trustworthy workers. Those who did not answer the gold question correctly and were below a certain threshold were classified as low quality workers [15]. We chose the top 10% tasks with the highest average accuracy for each task type to be the gold questions. Table 10 shows the relative proportions of low quality workers in the *CooVi* and *Baseline* groups under different types of tasks, with 70%, 80% and 90% of accuracy thresholds, respectively. Regardless of which threshold and task type we look at, the proportion of low quality workers in *CooVi* was lower than that in *Baseline*. This result also proves that, as we mentioned earlier, income sharing schemes reduce the attractiveness for cheaters.

In contrast, some workers spent an abnormally long time to complete HITs. If this happens in co-op conditions with shared rewards, workers who took much time to accomplish a few tasks may be classified as *slower-workers* or *free-riders*. Slower-workers are inefficient workers who conscientiously completed all HITs. Slower-workers are defined as having an average work accuracy and long HIT completion time. Rockart [43] defines free-riders as people who enjoy the

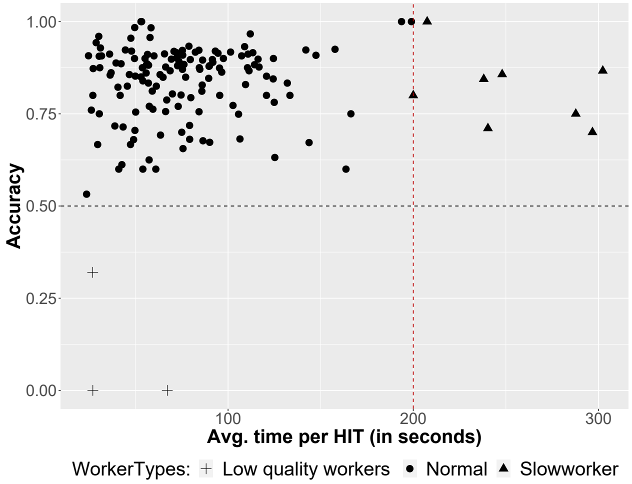


Fig. 4. Accuracy with respect to the average time per HIT for each worker in shared reward conditions.

benefits of commodities without contribution. In our case, free-riders are workers who leverage the reward scheme to gain a substantial reward without earnestly completing HITs. Free-riders' task quality is worse than random, so they are symbolised by having lower accuracy and long average completion time per HIT.

The presence of free-riders and slow-workers could harm the proposed reward allocation methods since the rewards received from the money pool is a function of the time spent on the tasks. Too many slow workers can ultimately negatively affect the income of others. The presence of free-riders is worse than slow-workers for the functioning of co-op schemes. Instead of sharing risks, they increase the risk for group members as incomes are distributed without any benefit. Therefore, we look at the presence and the effect of free-riders and slow workers on our co-op reward models.

To begin with, we look at the average HIT completion time per worker and identify free-riders and slow-workers in the *CooHi*, *CooHiSo* and *CooVi* conditions. Due to different task-inherent difficulty levels, it is hard to find an actual threshold value to identify what is an abnormally long task duration. However, looking at task time distributions, slowest workers in the group can be defined as having an average task completion time per HIT was found to be longer than two standard deviations (2SD) away from the mean value in the five control groups. Figure 4 represents the accuracy and average completion time per HIT of all workers in the conditions with shared rewards. We observed that the presence of few slow workers with normal accuracy, but we cannot find any evidence of the presence of free riders (i.e., bottom-right corner in Figure 4). Some workers with low accuracy did exist, but their average HIT working duration was quite short as well and thus it did not have a major impact on the reward of others in the group.

We found that the number of HITs finished by slow workers is only a small part of the total number of finished HITs (30 HITs, 0.4%) and slow workers in the *CooHi* and *CooVi* spent a minor proportion of the total time taken by all group members collectively (1.3% and 2.12%, respectively). Therefore, slow workers in the *HidIn* and *CooVi* conditions produced a negligible effect on the reward sharing scheme as the money that they received only consumed a tiny part of the group money pool.

Key Findings. We found that when platforms provide the same transparency of information, the reward-sharing group attracts fewer cheaters. We also found an absence of free-riding and the influence of slow workers on others in the group is minimal.

5.5 Post-Survey Analysis

To understand workers' perceptions of the co-op risk-sharing and reward-sharing strategies, we invited all participating workers to complete an optional survey at the end of the study. Workers were asked to indicate the extent to which they agreed with 4 different statements on a 5-point Likert scale ranging from *1: Strongly Disagree* to *5: Strongly Agree*. These statements are presented below:

- *CoopSlower* — I think that it is fair for someone slower than me in completing tasks (HITs) to get the same payment as me in a cooperative team.
- *CoopRandom* — I would feel comfortable to be in a cooperative team where risks and rewards are shared with random workers.
- *CoopKnown* — I would rather share risks and rewards with workers I know or workers who I am familiar with (for example, those people I converse with on forums).
- *CrowdUnion* — A crowd workers' union would be useful.

We collected responses from 74 different workers across the 5 conditions. Table 11 presents the results from our post-study survey. We found that workers tend to agree with the idea of the cooperative teams sharing risks and rewards (indicated by average scores > 3). Interestingly, we found that workers were in support of sharing their rewards with slower workers in the cooperative teams on average. This was particularly significant in conditions without shared rewards in the study; more than 93% of *Baseline* workers (*M=3.6*) and 71.4% of *HidIn* workers (*M=3.36*) indicated their support. Also, we found that workers preferred to participate in cooperative teams with strangers than with other familiar workers. At the $\alpha = 0.1$ level of significance, a Wilcoxon signed-rank test revealed a significant difference in this preference of workers to cooperate with strangers (Median=4) compared to familiar workers (Median=3);$z = -1.836, p = .066$. Finally, the respondents believed that a crowd workers' union would be useful, with over 81% of workers who reported a score of ≥ 3 on the Likert-scale. As mentioned earlier, a co-op reward scheme could also consider a share of the total reward not to be re-distributed among members but rather to be allocated to other mutually beneficial purposes (e.g., a union).

Table 11. Post-survey results about workers' perspective on CrowdCOOP (average on a 1-5 scale).

	Baseline	*HidIn*	*CooHi*	*CooHiSo*	*CooVi*
#Workers	15	14	9	17	19
CoopSlower	3.60	3.36	3.33	3.35	3.05
CoopRandom	3.73	3.14	4.00	3.65	3.68
CoopKnown	3.40	3.14	3.11	3.24	3.63
CrowdUnion	3.53	3.57	3.33	3.12	3.53

6 DISCUSSION

6.1 Risk Mitigation Tools Versus Reward-Sharing

Some tools are already available to help workers mitigating risks by avoiding bad tasks and requesters [36, 48]. However, utilizing such tools may lead to social divisions among workers [53] as those workers not using the tools may end up carrying most of the risks. The number of high-income tasks is limited, and workers who do not use risk mitigation tools see their exposure to high-income tasks reduced. Compared to these tools, reward sharing strategies do not create social divisions among members of the group as the proposed method encourages risk *sharing* rather than risk *avoidance*. Workers outside the co-op group would thus not be negatively affected (other than missing out on the risk mitigation benefit).

On the other hand, the proposed risk and reward sharing scheme encourages the coexistence with crowd work tools. As an example, Chiang et al. [10] introduced Crowd Coach: a Chrome extension that allows workers to provide suggestions on tasks they work on. Through these suggestions, other workers can improve their skills. However, workers under traditional piecework reward conditions may not be willing to provide such guidance as the time dedicated to this would be taken away from additional HIT completion. Thus, without additional financial incentives, it is not likely that such peer-support approach could gain in popularity. Our proposed reward scheme may be used to encourage workers to invest some time using mutual help tools given the shared income pool.

6.2 Piecework Versus Reward-Sharing

Piecework is defined as workers getting a fixed rate for each performed task [1]. In piecework, workers can achieve best performance by repeated "learning by doing" [33]. However, in some micro-task crowdsourcing platforms, the diversity of tasks makes it difficult for workers to maximize their efficiency, and the high cost of learning also makes workers reluctant to try new and more challenging tasks as it may lower their earnings [36]. By contrast, the proposed co-op reward scheme emboldens workers to take up advanced and longer HITs by sharing with others the associated financial and career risks.

Besides, piecework may discourage workers from sharing their insights on efficient task execution. If the average task completion time is reduced, requesters may reduce HIT rewards if they are set based on expected completion time [8]. Additionally, previous research show that profit-sharing incentives provide an effective way to encourage cooperation between individuals and improve effort and productivity [17]. Working in a cooperative may create an intrinsic motivation for workers as they help each other by sharing risks. Jakob et al. showed that intrinsic motivation can enhance labours' data quality in crowdsourced markets [44].

6.3 Guidelines for Crowdsourcing Cooperatives

Transparency is vital when establishing cooperative platforms. Reward sharing schemes may have no significant impact on the number of tasks completed by employees. Still, visible rewards and platform transparency play a crucial role in motivating employees to complete more work. Workers need to know how much money they can potentially earn for each task they complete and implicitly assign a value to the effort they are investing. In the long run, the impact of information transparency in the traditional compensation model is limited. However, information transparency still plays an essential role in reward-sharing schemes as income self-assessment is more complex and less obvious by design (i.e., it is not the simple sum of individual rewards).

Our post-study survey confirms that workers felt comfortable sharing risks and rewards, even in the presence of slow workers. We found that workers preferred to participate in co-op teams with strangers than with other familiar workers, so cooperatives may be easily formed through existing crowdsourcing communities (e.g., online MTurk worker forums). In a production-level deployment of the proposed approach, we envision a group of initial members to establish a collective where all members agree and abide to the spirit of mutual help. As the reward is shared among all members, they might all share organisation management responsibilities. Some participants may be elected to manage some aspects of the group (e.g., performing periodic quality assessment activities to identify free-riders) and be entitled to claim the time spent on such activities as time worth of reward allocation.

Our findings on MTurk may also be applicable to other crowd-work platforms such as Uber, Upwork, and TaskRabbit where cooperatives may be established independently of the platforms. Almost all workers in the gig-economy are independent contractors, and they suffer some of the same risks that we addressed in this work. For example, gig workers do not get any income when they are sick as they are not completing tasks. This risk is also one that can be shared among members of a cooperative. Workers in a co-op group may decide to allocate a part of the money pool (e.g., 1%) before the time-based redistribution happens, and use these funds to help members who are unable to work (akin to a sick leave).

7 LIMITATIONS AND FUTURE WORK

In this section we discuss the limitations that may arise from running the experiments under controlled conditions. First of all, this study did not employ quality control schemes, and all HITs were approved and rewarded. This allowed us to track good and bad quality work in the cooperative. Although slow workers and free-riders did not harm the cooperative group in our experiment, in a real deployment, their presence may lead to trust-related problems among workers. The presence of low quality workers may increase the task rejection rate and result in monetary loss for the group. Slow workers may limit the overall productivity and thus reduce the motivation of experienced workers. A peer-review system can be a possible strategy to control for slow workers and free-riders. Researchers in other fields have shown that members of unions have a great aversion to free-riders and are very active in identifying and eliminating them, even at a high cost [18]. Workers with abnormally short or long task completion time or with high rejection rates may be selected to go through an evaluation process led by experienced workers in the group. Detected inadequate workers may then be shepherded to improve, punished or removed from the cooperative group.

Another possible challenge for co-op reward schemes is the presence of workers who are multi-tasking. That is, workers who accept multiple tasks in parallel for the fear of missing out (as by accepting to work on a HIT, the HIT is allocated to them before the HIT batch runs out of work to be completed) and by doing so increase the computed task completion time which would then be used to calculate their reward. This issue could be addressed by considering the total work session time and the completed tasks within a session rather than the individual task completion times to re-allocate rewards as done in our experiments. Additionally, missing out on new HITs is not a risk that decreases workers' income under a co-op setting, as workers are paid based on the time they spent working on HITs independently of the number of HITs completed. Thus, while tools that aim to manage risks by estimating the HIT completion time (e.g., [46]) can be easily combined with the proposed reward sharing schemes, they may not be serving the same risk mitigation purpose but rather provide workers with time management functionalities (e.g., plan their work day activities).

In this experimental study, we used a fixed set of HITs available to participating workers. We believe that our findings concerning the proposed reward distribution models will hold in a dynamic marketplace with a continuous flow of newly available HITs. This is corroborated by workers who indicated a keen interest in participating in cooperative teams through the exit survey. Moreover, we have not studied the impact of the co-op group size across the different conditions. In future work, we will explore how team size, structure, and composition affects performances under the proposed reward distribution models.

Another limitation of the proposed approach is that experienced workers (who are known to be more efficient [27]) receive relatively small returns, especially for those who already know how to manage risks and perform tasks efficiently. Some risks, such as different inherent task difficulties in the same batch, cannot be easily avoided and are thus better shared within cooperative groups. However, the benefits they get from sharing risks could be lower than the benefits they get from being experienced, and thus decide that they are better off by not participating to co-op activities.

Therefore, reward strategies that take quality control into account will be an important element of crowd work cooperatives. One possible strategy to address this limitation is an approach were excellent and/or experienced workers receive an extra reward, which may come from lower rewards being allocated to slow workers or from an earmarked proportion of the money pool (see Section 6.3 where a similar approach has been suggested for sick leaves).

Some cultures believe that rest time should also be paid [35]. Research has shown that short breaks could improve workers' productivity and well-being [34]. Whether a short break will enhance the performance of crowdsourcing workers is a topic worth studying on its own [11]. In future work, we will study the impact of allowing each member of a reward sharing group to be absent for some time and having such a break being paid using a certain percentage of the money pool. We will also investigate how worker moods and engagement [56] within a CO-OP group are effected due to the sharing of risks and rewards.

8 CONCLUSIONS

In this paper, we have proposed and experimentally evaluated novel crowdsourcing risk and reward sharing schemes that enable crowd workers to move away from piecework payments and allow them to focus on performing crowd work with less stress and less time pressure. Our method also encourages workers to try diverse tasks. Even reward-motivated workers can take up challenging and complex HITs without worrying about missing financial rewards. To this end, we built a new crowdsourcing platform, CrowdCO-OP, by leveraging the API provided by MTurk, that allows us to perform controlled experiments comparing different reward distribution mechanisms and task information. Our findings show that co-op reward sharing creates a win-win situation for both workers and requesters: For workers, unrewarded activities are shared thus allowing for risk mitigation; For requesters, our alternative reward mechanisms lead to a more productive workforce and do not lead to lower quality data.

ACKNOWLEDGMENTS

This work is partially supported by the ARC Discovery Project (Grant No. DP190102141) and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328.

REFERENCES

[1] Ali Alkhatib, Michael S Bernstein, and Margaret Levi. 2017. Examining crowd work and gig work through the historical lens of piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4599–4616.

[2] Janine Berg. 2016. Income Security in the On-Demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers. *Comparative Labor Law & Policy Journal* 37 (2016), 543–576.

[3] Robin Brewer, Meredith Morris, and Anne Marie Piper. 2016. "Why would anybody do this?": Understanding Older Adults' Motivations and Challenges in Crowd Work. *Proceedings of the 2016 CHI Conference on human factors in computing systems* (2016), 2246–2257.

[4] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk:A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980

[5] Chris Callison-Burch. 2014. Crowd-Workers: Aggregating Information Across Turkers To Help Them Find Higher Paying Work. In *the Second AAAI Conference on Human Computation and Crowdsourcing*.

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

[7] Alessandro Checco, Jo Bates, and Gianluca Demartini. 2020. Adversarial Attacks on Crowdsourcing Quality Control. *Journal of Artificial Intelligence Research* 67 (2020), 375–408.

[8] Justin Cheng, Jaime Teevan, and Michael S Bernstein. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1365–1374.

[9] Iurii Chernushenko, Felix A Gers, Alexander Löser, and Alessandro Checco. 2018. Crowd-labeling fashion reviews with quality control. *arXiv preprint arXiv:1805.09648* (2018).

[10] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 37.

[11] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 628–638.

[12] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques and Assurance Actions. *CoRR* abs/1801.02546 (2018). arXiv:1801.02546 http://arxiv.org/abs/1801.02546

[13] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Florence, Italy, 238–247. https://doi.org/10.1145/2736277.2741685

[14] Serge Egelman, Ed H. Chi, and Steven Dow. 2014. *Crowdsourcing in HCI Research*. Springer New York, New York, NY, 267–289. https://doi.org/10.1007/978-1-4939-0378-8_11

[15] Carsten Eickhoff and Arjen Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16, 2 (2013), 121–137.

[16] Donna Harman Ellen M. Voorhees. 1999. Overview of the Eighth Text Retrieval Conference. In *the Eighth Text Retrieval Conference*. National Institute of Standards and Technology, Gaithersburg.

[17] S. ESTRIN and R. SHLOMOWITZ. 1988. INCOME SHARING, EMPLOYEE OWNERSHIP AND WORKER DEMOCRACY: Theory and Evidence. *Annals of Public and Cooperative Economics* 59, 1 (1988), 43–66.

[18] Ernst Fehr and Simon Gchter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90, 4 (2000), 980–994.

[19] Leon Festinger. 1954. A Theory of Social Comparison Processes. *Human Relations* 7, 2 (1954), 117–140. https://doi.org/10.1177/001872675400700202 arXiv:https://doi.org/10.1177/001872675400700202

[20] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–29.

[21] Ujwal Gadiraju and Gianluca Demartini. 2019. Understanding Worker Moods and Reactions to Rejection in Crowdsourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19)*. ACM, New York, NY, USA, 211–220. https://doi.org/10.1145/3342220.3343644

[22] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2018. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work: CSCW: An International Journal* (2018), 1–27. https://doi.org/10.1007/s10606-018-9336-y

[23] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14)*. ACM, New York, NY, USA, 218–223. https://doi.org/10.1145/2631775.2631819

[24] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.

[25] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 5–14.

[26] Snehalkumar Gaikwad, Durim Morina, Adam Ginzberg, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, Sharon Zhou, Mark Whiting, Karolina Ziulkoski, Alipta Ballav, Aaron Gilbee, Senadhipathige Niranga, Vibhor Sehgal, Jasmine Lin, Leonardy Kristianto, Angela Richmond-Fuller, Jeff Regino, Nalin Chhibber, Dinesh Majeti, Sachin Sharma, Kamila Mananova, Dinesh Dhakal, William Dai, Victoria Purynova, Samarth Sandeep, Varshine Chandrakanthan, Tejas Sarma, Sekandar Matin, Ahmed Nasser, Rohit Nistala, Alexander Stolzoff, Kristy Milland, Vinayak Mathur, Rajan Vaish, and Michael Bernstein. 2016. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. In *Proceedings of the 29th Annual Symposium on user interface software and technology (UIST '16)*. ACM, 625–637.

[27] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd Worker Strategies in Relevance Judgment Tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 241–249. https://doi.org/10.1145/3336191.3371857

[28] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *Proceedings of the Twelfth ACM International*

Conference on Web Search and Data Mining (WSDM '19). ACM, New York, NY, USA, 321–329. https://doi.org/10.1145/3289600.3291035

[29] Benjamin V. Hanrahan, David Martin, Jutta Willamowski, and John M. Carroll. 2019. Investigating the Amazon Mechanical Turk Market Through Tool Design. *Computer Supported Cooperative Work (CSCW)* 28, 5 (01 Sep 2019), 795–814. https://doi.org/10.1007/s10606-018-9312-6

[30] Karin Hansson and Thomas Ludwig. 2019. Crowd Dynamics: Conflicts, Contradictions, and Community in Crowdsourcing. *Computer Supported Cooperative Work (CSCW)* 28, 5 (01 Sep 2019), 791–794. https://doi.org/10.1007/s10606-018-9343-z

[31] 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*. ACM, Montreal QC, Canada, 1–14. https://doi.org/10.1145/3173574.3174023

[32] Ellie Harmon and M. Six Silberman. 2019. Rating Working Conditions on Digital Labor Platforms. *Computer Supported Cooperative Work (CSCW)* 28, 5 (01 Sep 2019), 911–960. https://doi.org/10.1007/s10606-018-9313-5

[33] Robert A Hart. 2016. the rise and fall of piecework. *IZA World of Labor* (2016).

[34] Robert A Henning, Pierre Jacques, George V Kissel, Anne B Sullivan, and Sabina M Alteras-Webb. 1997. Frequent short rest breaks from computer work: effects on productivity and well-being at two field sites. *Ergonomics* 40, 1 (1997), 78–91.

[35] Helene Jorgensen. 2002. Give Me a Break: The Extent of Paid Holidays and Vacation. *Washington, DC: Center for Economic and Policy Research.(September). Accessed April* 25 (2002), 2007.

[36] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P. Bigham. 2018. Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. In *HCOMP*.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (01 May 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[38] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (2017), 433–442.

[39] Xiao Ma, Lara Khansa, and Sung S. Kim. 2018. Active Community Participation and Crowdworking Turnover: A Longitudinal Model and Empirical Test of Three Mechanisms AU. *Journal of Management Information Systems* 35, 4 (2018), 1154–1187. https://doi.org/10.1080/07421222.2018.1523587

[40] 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*. ACM, Baltimore, Maryland, USA, 224–235. https://doi.org/10.1145/2531602.2531663

[41] 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

[42] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose, California, USA, 2271–2282. https://doi.org/10.1145/2858036.2858539

[43] Scott Rockart. 2016. *Free-Rider Problem, the.* Palgrave Macmillan UK, London, 1–3. https://doi.org/10.1057/978-1-349-94848-2_736-1

[44] 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. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[45] Jill Rubery. 2018. A GENDER LENS ON THE FUTURE OF WORK. *Journal of International Affairs* 72, 1 (2018), 91.

[46] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey P. Bigham. 2019. TurkScanner: Predicting the Hourly Wage of Microtasks. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3187–3193. https://doi.org/10.1145/3308558.3313716

[47] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1621–1630. https://doi.org/10.1145/2702123.2702508

[48] M. Six Silberman and Lily Irani. 2016. Operating an employer reputation system: lessons from Turkopticon, 2008-2015. *Comparative Labor Law & Policy Journal* 37, 3 (2016), 505–541.

[49] Ianna Sodré and Francisco Brasileiro. 2017. An Analysis of the Use of Qualifications on the Amazon Mechanical Turk Online Labor Market. *Computer Supported Cooperative Work (CSCW)* 26, 4 (01 Dec 2017), 837–872. https://doi.org/10.1007/s10606-017-9283-z

[50] Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation* 47, 1 (01 Mar 2013), 9–31. https://doi.org/10.1007/s10579-012-9176-1

[51] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. 2012. CrowdER: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment* 5, 11 (2012), 1483–1494. https://doi.org/10.14778/2350229.2350263

[52] Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 197–206.

[53] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article Article 24 (Nov. 2019), 28 pages. https://doi.org/10.1145/3359126

[54] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[55] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[56] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*. 373–382.

Received January 2020; revised June 2020; accepted July 2020