# Qrowdsmith: Enhancing paid microtask crowdsourcing with gamification and furtherance incentives

EDDY MADDALENA, University of Udine, Italy
LUIS-DANIEL IBÁÑEZ, University of Southampton, United Kingdom
NEAL REEVES, King's College London, United Kingdom
ELENA SIMPERL, King's College London, United Kingdom

Microtask crowdsourcing platforms are social intelligence systems in which volunteers, called crowdworkers, complete small, repetitive tasks in return for a small fee. Beyond payments, task requesters are considering non-monetary incentives such as points, badges and other gamified elements to increase performance and improve crowdworker experience. In this paper, we present Qrowdsmith, a platform for gamifying microtask crowdsourcing. To design the system, we explore empirically a range of gamified and financial incentives and analyse their impact on how efficient, effective, and reliable the results are. To maintain participation over time and save costs, we propose furtherance incentives, which are offered to crowdworkers to encourage additional contributions in addition to the fee agreed upfront. In a series of controlled experiments we find that while gamification can work as furtherance incentives, it impacts negatively on crowdworkers performance, both in terms of the quantity and quality of work, as compared to a baseline where they can continue to contribute voluntarily. Gamified incentives are also less effective than paid bonus equivalents. Our results contribute to the understanding of how best to encourage engagement in microtask crowdsourcing activities, and design better crowd intelligence systems.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: Qrowdsmith: Enhancing paid microtask crowdsourcing with gamification and furtherance incentives

## 1 INTRODUCTION

In recent years crowdsourcing has become established as a convenient way to scalably collect human-generated data annotations for many types of applications, such as surveys [12], entity linking [9], and urban mapping [35]. One of the most popular approaches to crowdsourcing is to break down the required work in *microtasks*: short, simple, self-contained units of work that can be given to human workers for them to solve as much as they can or want [24]. Microtasks are generally completed independently by many workers and then aggregated to facilitate quality assurance [4]. As a result, a large body of crowdsourcing research has considered methods for optimising worker engagement, retention and performance in a range of task types and contexts [22, 47, 59].

A key aspect for effective microtask crowdsourcing is the workers' motivation to perform the tasks. The most common motivation is financial, where workers are paid a fixed amount of money in exchange for the completion of a set of tasks [43]. Several commercial crowdsourcing platforms such as Prolific [1], Amazon Mechanical Turk

[1]Prolific website: https://prolific.co/

Authors' addresses: Eddy Maddalena, University of Udine, Udine, Italy, eddy.maddalena@uniud.it; Luis-Daniel Ibáñez, University of Southampton, Southampton, United Kingdom, l.d.ibanez@soton.ac.uk; Neal Reeves, King's College London, London, United Kingdom, neal.reeves@kcl.ac.uk; Elena Simperl, King's College London, London, United Kingdom, elena.simperl@kcl.ac.uk.


2157-6904/2023/6-ART $15.00
https://doi.org/10.1145/3604940

(mTurk)[2] or Appen[3] serve as intermediaries between task designers and workers and facilitate payments to workers.

The impact of financial rewards on user engagement has been a common focus of research in the domain of microtask crowdsourcing. A number of studies suggest a strong association between higher payments and engagement [38], although findings around the quality of contributions have been more mixed [33]. Conversely, more recent studies suggest that financial incentives alone are not enough to drive engagement and that tasks must appeal to workers' intrinsic motivations and sense of fairness to effectively mediate the effectiveness of increased pay [11, 31, 56]. There is increasing evidence to suggest that the implementation of non-monetary *furtherance incentives* is beneficial for the overall quantity and quality of work. Analyses of the impact of increased financial incentives on engagement have shown mixed results, suggesting workers are driven by other factors than simply monetary rewards alone [56]. Gamification mechanisms such as points and badges have been demonstrated to delay task abandonment and increase participation among paid crowdworkers [19]. Other research suggests that despite acting as extrinsic motivators, these mechanisms act distinctly from financial rewards by appealing to participants' intrinsic motivations to display their reputation and capabilities to others [16].

In this paper we introduce Qrowdsmith, a platform for gamifying microtask crowdsourcing. Qrowdsmith implements a configurable point system, bonuses and level scales, user profiles, badges, a global scoreboard and the granting of "special powers". It can interface with existing paid microtask crowdsourcing platforms to complement gamification with monetary incentives. Qrowdsmith is a crowd intelligence system, which draws its superior, "intelligent" performance from a mix of empirically designed system features and the combined knowledge and expertise of the crowdworkers it recruits and rewards. A critical consideration for systems like Qrowdsmith to work in practice is to understand how crowdworkers react to different incentives and combinations thereof. This is achieved through a combination of theory in sociotechnical systems, motivations, and incentives, which helps us identify desirable incentives that the system supports, and empirical experimentation, in which we observe crowd behaviour and turn it into actionable knowledge [54] and iterations of the system design.

Microtask crowdsourcing research has focused on studying monetary incentives or specific gamification components like contests [18]. Much fewer papers have tackled the problem of identifying what incentives work best as *furtherance*, *i.e.*, that encourage voluntary additional contributions. Using Qrowdsmith, we study the gamification incentives offered as *furtherance* over monetary ones impact the quality and amount of work paid workers do. We conduct experiments on an image annotation task assigning workers to one of eight experimental conditions varying based on number of rounds (1 or 11), incentive type (gamified, financial or none) and payment (£0.20 or £0.50 per task). We ask the following research questions:

- RQ1: Can furtherance incentives prompt workers to undertake more work? What are the most effective furtherance incentives?
- RQ2: How furtherance incentive affects the speed at which workers perform tasks?
- RQ3: How furtherance incentives affect the quality of microtask work?
- RQ4: How furtherance incentives affect the reliability of crowdsourced data
- RQ5: What combination of incentives is more cost-effective with respect to work quantity? What with respect to work quality?

We analyse the number of tasks workers performed prior to, during and after the offer of furtherance incentives, as well as the accuracy of – and inter-annotator reliability measures.

[2]Amazon Mechanical Turk website: https://www.mturk.com/
[3]Appen website: https://appen.com/

## 2 BACKGROUND AND RELATED WORK

### 2.1 Motivation and incentives in paid microtask crowdsourcing

Financial incentives and monetary rewards represent the main source of motivation in paid crowdsourcing platforms [27, 43]. Mechanical Turk workers from both India and the US identified the opportunity to make money as the most significant factor driving their use of the platform [34]. In a survey of motivations conducted by Martin et al., the authors concluded that financial incentives exceed all other motivations – that is, workers will prioritise otherwise undesirable tasks that pay highly over fun or engaging tasks that pay poorly [37]. Unfair pay from requesters has been demonstrated to significantly encourage workers to abandon tasks and is the factor most associated with worker dissatisfaction [5]. Similarly, an analysis of task abandonment rates by Han et al., found that higher financial incentives were associated with lower rates of task abandonment among workers [23].

This is not to suggest that non-monetary incentives do not influence crowdworker engagement. A systematic review of motivations in crowdsourcing tasks conducted by Spindeldreher et al., found that task enjoyment had the strongest positive effect on a worker's decision to participate in a crowdsourcing task [52], although it should be noted that this analysis considered types of crowdsourcing beyond paid microtasks. Workers' individual interests, skills and perceived personal capacity to complete tasks all play significant roles in the types of task they choose, particularly where pay is equal [41]. Furthermore, in addition to payment, workers will select their tasks based on a desire for diversity [45].

Further studies have suggested social, collaborative and community motivations. A study of participation in microtask crowdsourcing found personal growth and the opportunity to contribute to a community as key drivers for initial participation [10]. Awarding points and other non-monetary incentives has been shown to enhance participation in crowdsourcing activities by contributing to participants' impressions of fairness and recognition [55]. Experiments with collaborative, paired models of contribution have demonstrated that crowdworkers are also influenced by social incentives, responding not only to financial incentives for themselves but also for their paired partner [17]. Subsequent research has demonstrated that participation in paid crowdsourcing can be equally increased using a social, conversationally-driven task interface [47].

### 2.2 Gamification to drive engagement

Gamification has been used heavily within citizen science and to drive volunteer participation in research, particularly in so-called 'Games With A Purpose' which embed tasks within game-driven activities [49]. Analyses of the effectiveness of competitions in such projects suggests that participation is heavily driven by a desire to match and exceed the performance of fellow participants in terms of point-scores and leaderboard features [30, 44]. Nevertheless, the effectiveness of inter-participant competition is significantly influenced by the degree to which high performance is seen as achievable, with participants tending to display *reduced* levels of engagement where fellow volunteers' scores are seen as excessive [15, 30].

There is some disagreement within the literature as to the effectiveness of collaborative gamification mechanisms relative to or in place of competition. A study of competition participation in the Game With A Purpose *EyeWire* found collaborative gamified contests to lead to greater levels of participation than competitive equivalents [48]. Conversely, an analysis of the impact of gamified elements on participation in enterprise crowdsourcing conducted by [1] found competitive elements such as leaderboards were associated with significantly greater levels and increases in participation relative to collaborative mechanisms. Results from a field experiment by Morschheuser et al., argue that a team-based competition structure has the greatest impact on both engagement and motivation, combining both collaborative and competitive elements to maximise participation [40].

Gamification elements also serve a key role as *reward* mechanisms, through features such as badges and achievements. These reward mechanisms have been suggested to increase the quantity of submissions in crowdsourcing activities, meeting volunteer motivations through the recognition of their competencies [21]. In this sense, these incentives serve a similar role to financial payments, although this should not suggest that such incentives can *replace* monetary rewards.

Gamification has also been associated with increased worker retention. The opportunity to create and maintain worker avatars and profiles has been demonstrated to be associated with increased user retention and reduced perceptions of cognitive load [46]. Collaborative or social goals are associated with longer user retention than individual user workflows [17]. However, where these rewards are based on static milestones, there is a risk that they fail to encourage user retention, instead encouraging users to leave upon achieving a given reward [28]. More broadly, if these gamified rewards are to be effective, they must be carefully implemented to avoid feeling irrelevant or undesirable [26].

A comparative analysis of gamification methods for microtask crowdsourcing platforms found that tasks with gamified elements received significantly more submissions from workers than non-gamified tasks, but did not have a significant effect on the quality of submissions [32]. This was particularly true where bottom-up gamification was employed, allowing workers to choose the gamification features that they wished to see. However, a systematic literature review conducted by Morschheuser et al. conversely suggested that gamification-driven incentives result are associated with greater numbers of contributions and accuracy of submissions than financial incentives alone [39].

We build on this prior literature by explorign the impact of gamified elements on both the quantity and quality of participant contributions. In addition to more traditional reward elements such as badges and achievements, we explore the role of Qrowdsmith-specific functionality rewards designed to appeal more to intrinsic motivations such as enjoyment. We also make use of profile and avatar functions and perform comparative analyses of distinct gamification strategies to further explore how the relevance and desirability of gamification strategies impacts engagement.

### 2.3 Furtherance Incentives

The question of worker retention has long been of concern in crowdsourcing research and studies have attempted to address this through a range of approaches with differing levels of success. Most commonly, these incentives have taken the form of financial incentives and bonus payments. Generally studies have shown that the larger these bonus incentives, the more likely participants are to take part in subsequent activities [3, 53]. This increased worker retention plays a significant role in task completion rates in crowdsourcing platforms, potentially higher than the role of worker numbers and new worker registrations [20].

However, there is increasing evidence to suggest that the effectiveness of these strategies is dependent on the payment schedule and requirements introduced to achieve this pay. Difallah et al., demonstrated that the most successful strategies for encouraging worker retention were a milestone-based strategy where workers earned bonus rewards if they achieved specific goals [13]. Nevertheless conflicting findings suggest that while such milestone strategies are effective, workers tend to work for longer and complete more tasks without explicit completion goals than with them [51].

In addition to these financial incentives, experiments have also been carried out with gamified or social incentives designed to encourage worker retention. Feyisetan and Simperl experimented with a number of gamified furtherance incentive types, such as points, leaderboards and feedback functions, finding that these incentives successfully drove workers to contribute for longer and more accurately than financial incentives alone [19]. In a subsequent study, the authors explored the role of social pressures and collaboration on worker retention and found that these social incentives similarly encouraged workers to contribute for longer [17].

| Source | Finding | Experiment Design |
|---|---|---|
| [37] | Workers prioritise high paying tasks | Survey (N-794)<br>System log analysis (N=28466) |
| [5] | Unfair pay associated with higher levels of worker dissatisfaction | Online experiment (N=513) |
| [5, 23] | Higher pay associated with reduced rate of task abandonment | Online experiment (N=513, N=100) |
| [52] | Task enjoyment has strongest impact on decision to engage with task | Systematic literature review |
| [45] | Worker task selection influenced by desire for diversity of task type | Online experiment (N=23) |
| [55] | Points associated with feeling of fairness and recognition | Survey (N=235) |
| [30, 44] | Participation driven by competition on basis of point score and leaderboards | Online experiment (N=120)<br>Survey (N=235) |
| [15, 30] | Excessive competitor achievements lower drive to participate | Survey (N=545)<br>Interview (N=18)<br>Online experiment (N=120) |
| [48] | Collaborative challenges more engaging than competitor equivalents | System log analysis (N=10,296) |
| [1] | Competition more engaging than collaboration | Online experiment (N=101) |
| [46] | Gamification rewards associated with worker retention | Online experiment (N=800) |
| [32] | Gamification rewards associated with increased quantity but not quality of contributions | Online experiment (N=106) |
| [39] | Gamification rewards associated with increased quality and quantity of contributions compared with financial rewards | Online experiment (N=459) |
| [3, 53] | Larger bonus payments associated with increased worker retention | Online experiment (N=359, N=331) |
| [28] | Achievement of fixed milestone goals encourages task abandonment in workers | Online experiment (N=13,000) |
| [13] | Milestone goals most successful strategy for encouraging worker retention | Survey (N=40,000) |
| [51] | Workers retained for longer when milestone goal strategy not employed | Online experiment (N=602) |

Table 1. Comparison table showing main findings from wider literature.



Similar results have been seen in more recent work using social, conversational interfaces within microtask crowdsourcing platforms [47]. Finally, Law et al. explored the role of intrinsic motivations and curiosity on workers, finding that these incentives successfully encourage workers to contribute for longer, although payment was also noted to be a significant driver for further participation [31].

Building on these prior findings, we use distinct financial incentive levels including both a lower paid and higher paid level. We explore the effectiveness of financial bonuses relative to non-bonus and gamified incentive strategies. Our experiment includes opportunities to contribute with and without static milestones as a means of comparison for both incentivisation strategies. A comparison of key findings, as well as the source and methodology used to generate these findings can be found in Table 1.

## 3 QROWDSMITH

In this section we describe the Qrowdsmith platform, its functionalities, and how we recruited workers from popular paid microtask crowdsourcing platforms. First, we define terms recurrently used in this paper.

### 3.1 Terminology

- **Worker** or **crowdworker**: a human registered to a crowdsourcing marketplace, s.a. `Prolific`, `mTurk` or Appen , willing to perform crowdsourcing tasks in exchange for payment;
- **Human Intelligence Task (HIT)** (or **HITs for plural**, shortened as *task*): a short task, typically designed to be completed in a few minutes, published in a crowdsourcing marketplace. A HIT is comprised of one or more *rounds*.
- **Round**: the minimal unit of work of a HIT. For example, for a HIT that asks workers to label the sentiment of $N$ tweets, a *round* is the labelling of a single tweet.
- **Study**: the name the `Prolific` platform gives to a HIT;
- **User**: a worker that is registered in Qrowdsmith;
- **Furtherance incentives (FI)** or *rewards*: offers that Qrowdsmith makes to its users when they are about to leave the task to attempt to keep them engaged;
- **Task abandonment**: when a user leaves Qrowdsmith before completing all the rounds required to get paid.
- **FI abandonment**: when a user who accepted a furtherance incentive to continue working leaves without completing the number of rounds required to get the incentive.

### 3.2 Gamification components

In this section we present the gamification components implemented in Qrowdsmith. Figure 1 shows a screenshot of the Qrowdsmith interface with each of the six numbered circles corresponding to a gamification component as described below.

(1) **Profile and Levels**. Qrowdsmith users can customize their profile by changing their nickname, avatar, distinctive color, and country flag. Figure 1 (1) shows the user profile card as seen by the user. The avatar can be chosen from a pre-defined list of eight images. The user profile shows the current user level. Levels are awarded to users upon reaching specific score thresholds, as shown in Table 2. These values were inspired by Smahel et al. [50] who suggest levelling systems should work by making it easier for users to level up at the beginning of the game and then slowly increasing the difficulty or time required to reach higher levels. Our levelling system uses the Fibonacci series (starting from number three), similar to [2].

(2) **Chat**. Qrowdsmith offers a general public chat where logged-in users can communicate with each other in real-time. The chat is always available by default, but an administrator can disable it for a certain task, e.g. to avoid the sharing of solutions. Figure 1(2) shows the chat panel.

(3) **Bonuses** Qrowdsmith point rewards triggered by an event. Contrary to badges and levels, users can get the same bonus multiple times. Qrowdsmith implements two types of bonus trigger events:
   - **N rounds** Triggered when a user completes $N$ rounds., with $N$ a configurable value.
   - **Random** Triggered randomly, with a configurable probability, after the completion of a round. The number of bonus points is also randomly chosen from an integer value between 5 and 10.

(4) **Badges**. Badges are distinctive emblems that users can collect and display in their user profiles. Qrowdsmith implements two types of badge:
   - **Point-based badges**: similar to levels, these badges are awarded to users upon achievement of a predefined score threshold (see Table 2).

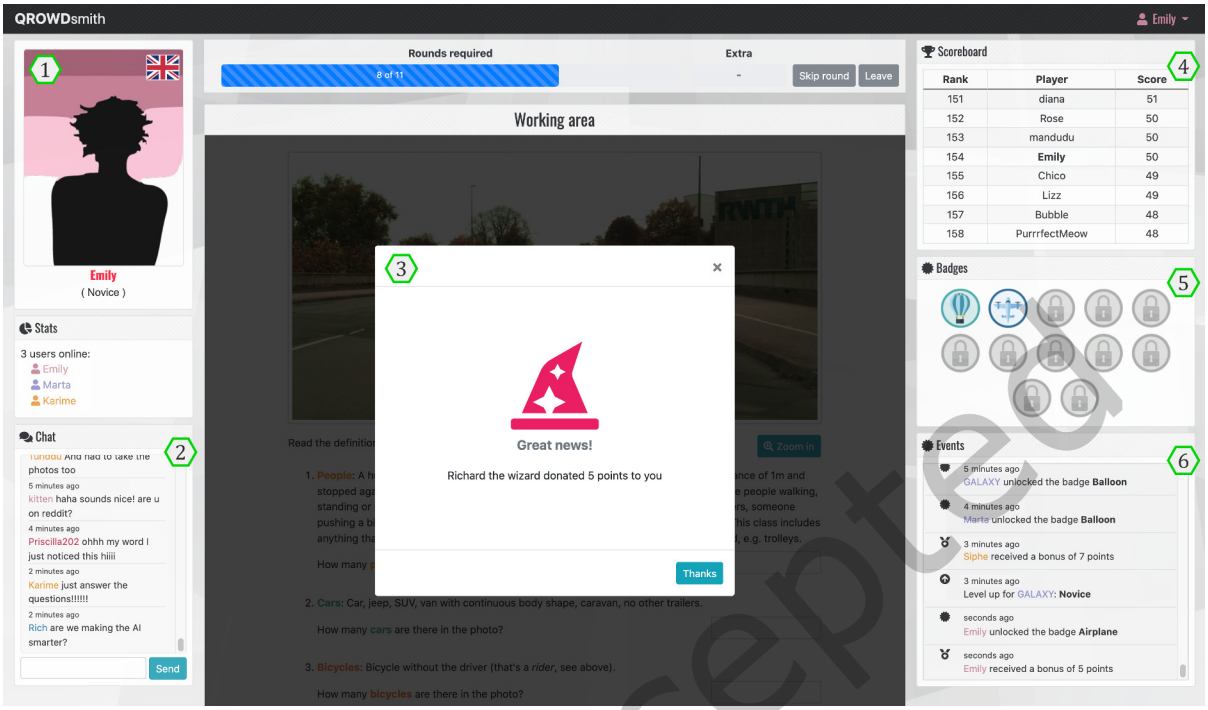



Fig. 1. Screenshot of the Qrowdsmith main menu, showing the six gamified components implemented on Qrowdsmith to increase user engagement: (1) Profile and levels, (2) Chat, (3) Bonuses, (4) Scoreboard, (5) Badges, and (6) Events.

- **Furtherance badge**: a single "Special" badge awarded to users that accept and complete an extra number of rounds.

(5) **Scoreboard**: shows the current user score, their position in the overall point ranking and the scores and positions of users immediately below and above in the ranking. In general, the user is shown the three users immediately above and the four users immediately below them. Users in the top 3 of the scoreboard see the top 8 users; users in the bottom 4 of the ranking see the bottom 8. See Figure 1(3). Each completed round awards one point. When a user earns points, a sound effect and a visual effect around the leaderboard are triggered.

(6) **Events**: Qrowdsmith notifies users in real-time about relevant events that take place within the platform (Figure 1(5)). The events shown are:
- another user receives a bonus
- another user achieves a new level or unlocks a new badge;
- a new user registers on Qrowdsmith;

All gamification components can be globally disabled. In this mode, Qrowdsmith is equivalent to a traditional crowdsourcing platform.

## 3.3 Task lifecycle

Tasks in Qrowdsmith follow the general workflow depicted in Figure 2. First, the user is shown the task instructions and the number of mandatory rounds $m$ required to complete the task. Users can abandon the task at any moment

| Type | Name | Score |
|---|---|---|
| Level | Newbie | - |
| Badge | Balloon | 3 |
| Level | Novice | 5 |
| Badge | Airplane | 8 |
| Level | Competent | 13 |
| Badge | Rocket | 21 |
| Level | Master | 34 |
| Badge | Bronze Cup | 55 |
| Level | Champion | 89 |
| Badge | Silver Cup | 144 |
| Level | Maestro | 233 |
| Badge | Gold Cup | 377 |
| Level | Commander | 610 |
| Badge | Platinum Cup | 987 |
| Level | Grand Duke | 1597 |
| Badge | Qrowdsmith | 2584 |
| Level | Qrowdsmith | 4181 |

Table 2. Qrowdsmith's levelling progression.



by clicking on the "Leave" button and confirming their intention on a pop-up dialogue. If a user abandons the task before completing *m* rounds they will not be paid. Users who completed all mandatory rounds may continue working on any number of additional rounds without payment. We refer to these rounds as *extra* or *voluntary*. If a user leave after completing *m* rounds and the task has furtherance incentives enabled, they are offered a furtherance incentive to perform *n* additional rounds. If the user does not accept the furtherance incentive, the task ends and the user gets paid. If the user accepts the furtherance incentive, they are taken back to the main screen to continue the task. We refer to the rounds performed towards the achievement of the incentive as *furthered*. If the user leaves before completing the *n* further rounds, they don't get the furtherance incentive, the task ends and they get paid. If they complete the *n* rounds, they get the furtherance incentive. They can either leave immediately and get paid, or keep working without further incentive for as long as they want. We refer to these type of rounds as *extra after furtherenace incentive*. Users are not offered a second furtherance incentive during the same task.

### 3.4 Task interface

When a user chooses an available task from the main menu, they are first shown the task description and instructions. The user can begin the task by clicking the "Start" button, leading to the task interface (Figure 3). The task interface is divided into two sections. First, the panel at the top of Figure 3 includes a bar showing the progress towards the required number mandatory rounds and the number of extra rounds. The panel in the Figure shows a user that complete 11 out of 11 mandatory rounds and 3 additional rounds. This panel includes a "skip round" button that allows users to get another round. A skipped round is not counted on the progress bar and does not negatively impact the user. Finally, the advancement panel includes the "Leave' button to leave the task after confirmation on a pop-up dialog box. The second section of the task interface is the working area (bottom of Figure 3), where the data of the current round and user controls are located. The example in the Figure

shows an image annotation task with the image to annotate and three input fields for users to introduce their annotations.

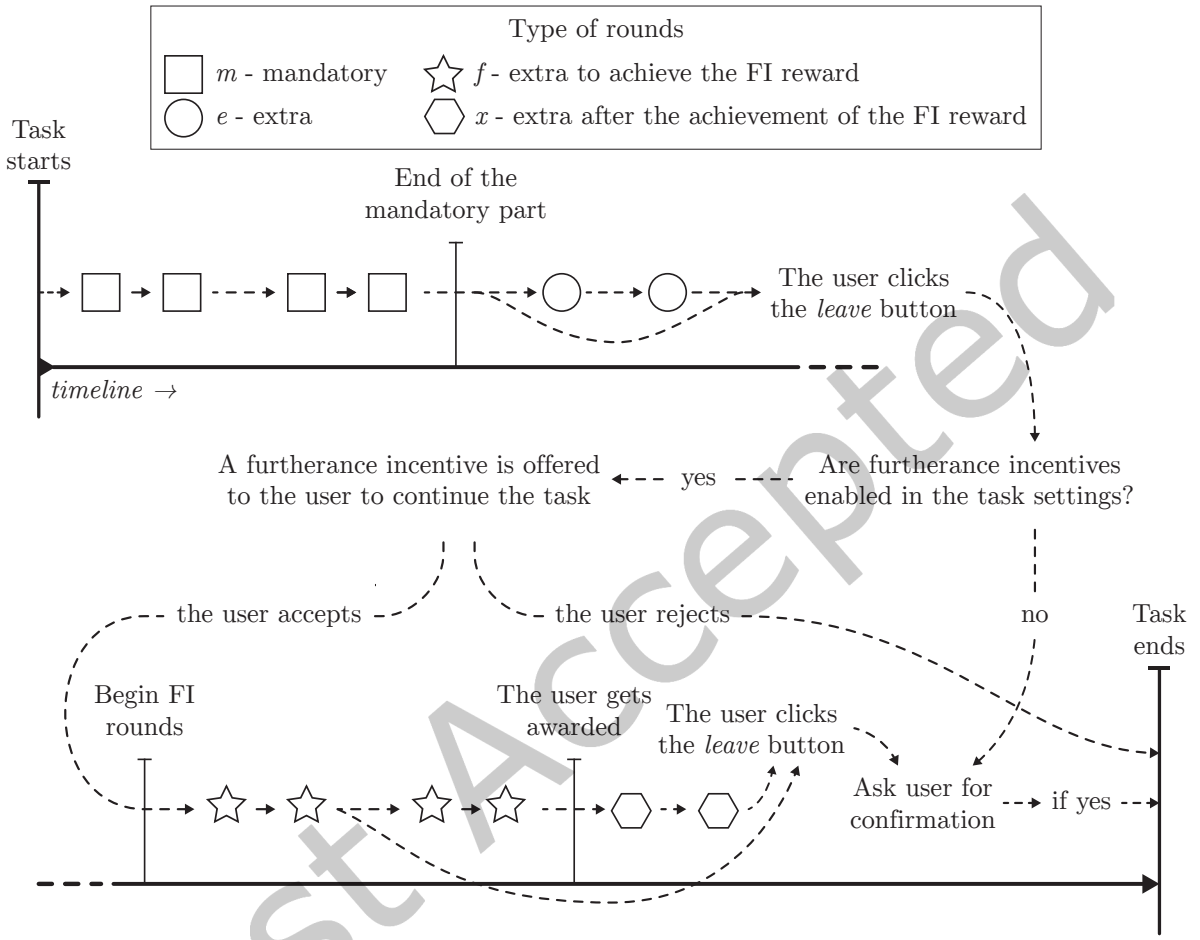



Fig. 2. Task workflow in Qrowdsmith. A sequence of rounds of type *Mandatory (m)*, *extra (e)*, *extra to achieve the Furtherance Incentive reward (f)*, *extra after achievement of the Furtherance Incentive reward (x)*



## 4 METHODS

### 4.1 Task

We designed an image annotation task where users to count the number of items featured in images. Each round, the user is shown a single image of an urban landscapes featuring at least two instances of either bicycles, buses, cars, motorcycles, trucks, persons and riders[4]. The user is asked to count the number of featured items of three

[4]Full descriptions for each item type are presented on this page https://www.cityscapes-dataset.com/dataset-overview/

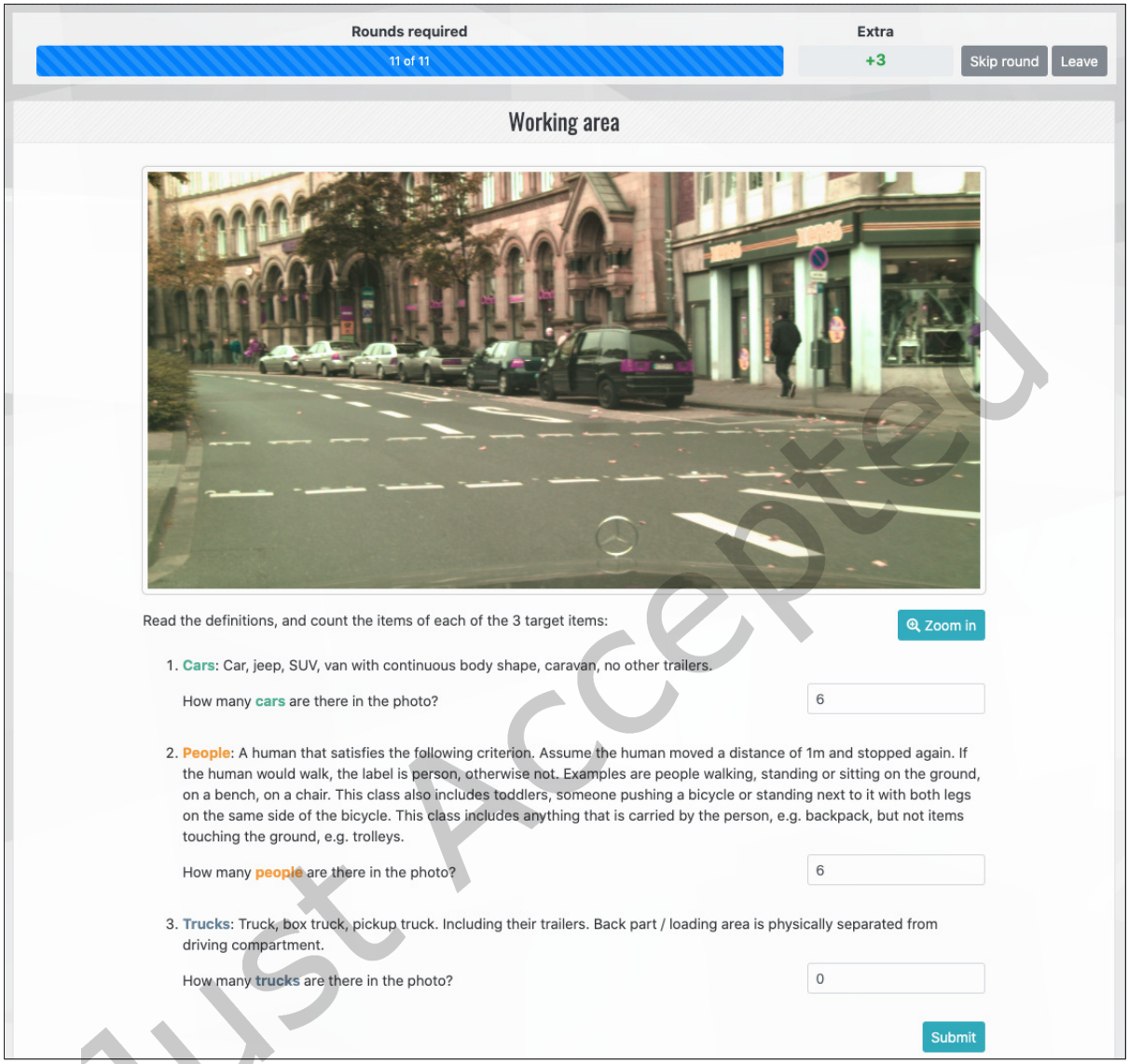



Fig. 3. The Qrowdsmith working area showing a round from an urban photo annotation task. The panel at the top tracks the user's progress: the progress bar tracks the completion of mandatory rounds (in the Figure, the user has completed all eleven mandatory rounds). The counter shows the number of extra rounds performed (three in the figure). The panel below shows the photo to be annotated with the input fields and their descriptions.

of the types and input them in a form below the image. Figure 3 shows an example of the round, including the progress bar and the counter of extra rounds.

### 4.2 Data

We used the *Cityscapes Dataset* [5] [7]. Cityscapes is comprised of 25 thousand images of urban landscapes from 50 cities annotated at different levels of granularity with polygonal annotation entities, including the number of entities of each of the types we defined for our task. We used the annotations in the dataset as a ground truth for our experiments. We used the *gtFine_trainvaltest* subset, that has 5000 images. From this initial dataset we filtered images that included at least one entity of two different item types described in section 4.1. To keep the complexity of the task low, we filtered out images with more than 10 occurrences of the same target item. The resulting dataset has 1014 pictures.

To assign the types to ask for each picture we randomly select three of the seven types, but repeating the draw until at least two of the three types has at least one instance in the picture. To ensure that we had enough annotations to calculate inter-agreement measures, the order in which images were presented to all participants was fixed.

### 4.3 Recruitment

Qrowdsmith's current implementation does not support monetary payment to users. For our experiments, we used `Prolific` as a proxy to recruit and pay workers. We used `Prolific` as it is the platform suggested by our institution's ethics policy. Note that Qrowdsmith's design is standalone, allowing the potential use of any other paid crowdsourcing platform.

To recruit workers, we created for each task a `Prolific` study describing the work to be undertaken, the payment and a declaration informing the task was part of a scientific study. Workers who agreed to participate were then redirected to Qrowdsmith. If the recruited worker was not registered in Qrowdsmith, they were invited to create a new user profile including unique username, country flag and avatar. Next, they were asked to read the *Information sheet for participants* and explicitly accept the terms of the *consent form* of the study specified on the Ethical Approval received by our study (released on 21/04/2021 from the King's College London Research Ethics Committee, number MRA-20/21-23018).

### 4.4 Experimental configurations

We designed eight experiment configurations described below and summarized in Table 3.

(1) **0.1**, a task of one round without any gamification or furtherance incentives enabled. Payment set to £0.20 upon completion.
(2) **0.11**: same as `0.1` but comprised of 11 rounds instead of 1. Payment set to £0.50 upon completion. The extra payment w.r.t. `0.1` is proportional to the extra amount of time needed to complete the task as estimated from a pilot study and assuming the time to read instructions and start the task is the same for both configurations.
(3) **1.1**: a single round task with all gamification components enabled but no furtherance incentives offered. Payment set to £0.20 upon completion, same as **0.1**
(4) **1.11**: same as 1.1 but comprised of 11 rounds instead of 1. Payment set £0.50 upon completion, same as **0.11**
(5) **Furtherance incentives**: 11 rounds paid at £0.50 upon completion plus the option of completing 11 additional rounds in exchange of a furtherance incentive. We considered four types of furtherance incentives:
   (a) **Special Badge (2.11 SB)**: a "Special" badge different from those detailed in Table 2. Once the badge is awarded, it is shown in the badge panel even after the end of the task.
   (b) **Qrowdsmith Points (2.11 QP)**: a fixed number of points, set to 11 in our experiments. Points are cumulative across tasks.

[5]The webpage of the Cityscapes Dataset: https://www.cityscapes-dataset.com/

| Exp. | RR | Gam | Reward type | ER | Payment (£) |
|---|---|---|---|---|---|
| 0.1 | 1 | - | - | - | .2 |
| 0.11 | 11 | - | - | - | .5 |
| 1.1 | 1 | ✓ | - | - | .2 |
| 1.11 | 11 | ✓ | - | - | .5 |
| 2.11 SB | 11 | ✓ | Special Badge | 11 | .5 |
| 2.11 QP | 11 | ✓ | Qrowdsmith Points | 11 | .5 |
| 2.11 SP | 11 | ✓ | Special Power | 11 | .5 |
| 2.11 MR | 11 | ✓ | Monetary Reward | 11 | .5 + .5 bonus |

Table 3. Summary of the eight experiment configurations; *Exp* is the experiment setting label, *RR* is the number of mandatory required rounds, *Gam* shows if gamification is enabled, *Further incentive* is the furtherance incentive, *ER* is the number of extra rounds necessary to get the incentive, and *Payment (£)* is the amount in British Pounds paid to users who completed the task.

(c) **Special Power (2.11 SP)**: the ability to see answers by another user, randomly chosen by the system, who has previously completed the round. The ability persists until the user leaves the task;
(d) **Monetary Reward (2.11 MR)** an additional monetary payment of £0.50.

### 4.5 Amount of work and completion time

For each participating user we counted the number of extra rounds performed voluntarily after completing the mandatory rounds, the number of rounds after accepting the incentive, and the number of extra voluntary rounds after completing the rounds needed to get the incentive.

To assess if there is a relationship accepting the furtherance incentive and the time to complete a task, we instrumentalised Qrowdsmith to measure the time in seconds taken by each participating user to complete each of the rounds they submitted, then, we compute the mean completion time for each user for each type of round and compare across the different experimental configurations. We don't consider the time to choose a task and read the instructions.

### 4.6 Attractiveness measures

To measure the attractiveness of a furtherance incentive *FI* we count the number of times *FI* is offered, the number of times *FI* is accepted and the number of times *FI* is completed and compute the:

- Accepted/Offered (A/O) ratio as number of acceptances divided by number of offers. Indicator of the FI's initial attractiveness to users.
- Completed/Accepted (C/A) ratio, as number of completed divided by number of acceptances. Indicator of level of user commitment to achieve the incentive.
- Completed/Offered (C/O) ratio, as number of completed divided by number of accepted. Indicator of FI's overall attractiveness.

Intuitively, a high value of each of the three ratios would mean a highly attractive incentive and viceversa. A high A/O ratio with low C/A ratio suggests an incentive that is attractive when offered, but not enough to keep users's interest all the way to completion. The opposite situation, low A/O ratio with high C/A ratio indicates an incentive that attract only a few users, but with enough commitment to do all the required work to achieve it.

## 4.7 Quality measures

To compare the impact on annotation quality among the experimental configurations we use the following measures:

*Absolute error:* given an image $i$ and a type $t \in T = \{bicycle, car, bus, truck, motorcycle, person, rider\}$, let $c_{i,t}$ be the number of items of type $t$ featured in image $i$ according to the ground truth. Given an user $u$, we refer to their annotation of an image $i$ as $i_u = (a_{t1}, a_{t2}, a_{t3})$ with $t1, t2, t3 \in T$ the types asked to the user in the task. We use $i_u(t1)$ to denote the value of the user's annotation $i_u$ for type $t1$ ($a_{t1}$ in our example). Then, the **absolute error** of a user $u$'s annotation of type $t$ in image $i$ with respect to the gold standard is defined as:

$$err_{u,t,i} = |c_{i,t} - i_u(t)| \tag{1}$$

For example, if an user counts seven cars in an image that according to the ground truth has nine cars, the absolute error is two.

*Annotation quality.* : the quality of an user annotation of image $i$ is defined as:

$$q_{i,t,u} = \begin{cases} 1 & \text{if } c_{i,t} = 0 \wedge i_u(t) = 0 \\ \frac{err_{u,t,i}}{\sqrt{(c_{i,t}+i_u(t))}+err_{u,t,i}+1} & \text{otherwise} \end{cases} \tag{2}$$

Our measure has the following three properties:

(1) the domain is in [0,1];
(2) the quality is 1 if and only if both the user annotation is equal to the ground truth;
(3) The impact of absolute error in quality is higher when the ground truth values are lower. Figure 4 shows the formula output for user and Citiscapes annotations in the range [0-10], notice that an absolute error of 2 when the true values are between 1 and 3 produces a quality of 0.6, whilst the same error when the true values are between 7 and 10 result in a higher quality of 0.71.

*Image annotation quality.* Let $T_{i,u}$ be the set of three annotations of image $i$ by user $u$. We define the image annotation quality of image $i$ by user $u$, $q_{i,u}$, as the mean quality of the annotations made by user $u$ for the image $i$:

$$q_{i,u} = \frac{1}{|T_{i,u}|} * \sum_{t \in T_{i,u}} q_{i,t,u}$$

*User quality.* Let $A_u$ be the set of image annotations provided by user $u$. We define the user quality $q_u$ as the mean of their image annotation qualities:

$$q_u = \frac{1}{|A_u|} * \sum_{i \in A_u} q_{i,u}$$

## 4.8 Reliability

Inter-annotator agreement is a widely used metric in crowdsourcing to measure how contributors agree in their annotations and estimate the reliability of the collected data [6, 36]. To assess the effect of furtherance incentives on reliability, we use *Krippendorff's alpha* to measure inter-annotator agreement across the four rounds types (mandatory, extra voluntary, extra to achieve incentive reward, extra voluntary after incentive) and the eight experimental settings (detailed in Table 3 )[29]. The output of Krippendorf's alpha is in the $[-1, 1]$ range where 1 indicates perfect reliability, a positive value indicates some reliability, zero indicates no reliability, and negative values indicates systematic disagreement. Note that Krippendorf's alpha allows, to a certain extent, for missing

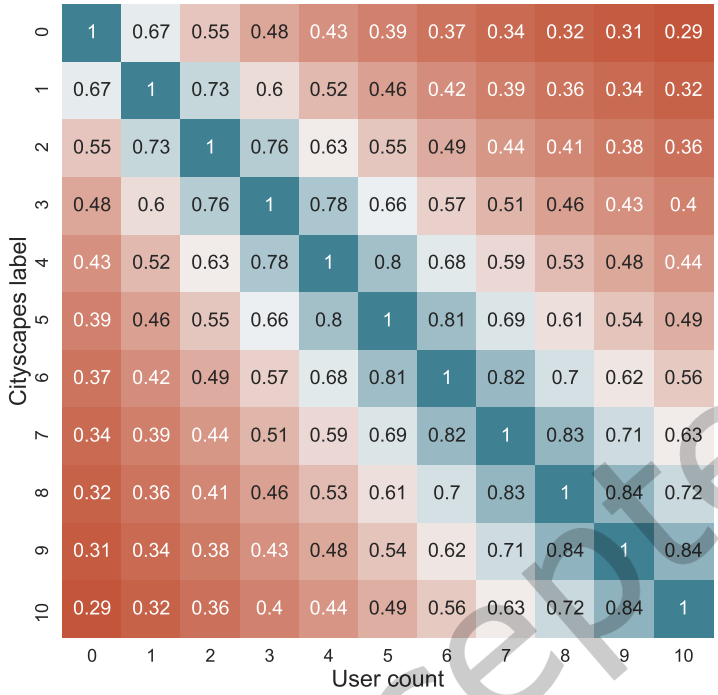



Fig. 4. Matrix of the quality scores given a Citiscapes label and a User count.s

annotations for some items, as may occur in our settings when an user abandons a task before completion, or refused a furtherance incentive.

## 5 RESULTS

In this section we present the results of the eight experiments listed in Table 3. All experiments were executed on a two week period between September and October 2021. To avoid the overloading the Qrowdsmith platform we limited the number of concurrent users to 10. We left each experimental setting task open until we reached 100 task completions, for a total of 800 unique users. *i.e.*, we did not allow users to participate more than once in the study.

The full cost of the experiment amounted to £474.81: configurations `0.1` and `1.1` costed £20 each, while the other six configurations costed £50 each. 34 users accepted and completed the monetary payment furtherance incentive, costing £17 total. `Prolific`'s service fee was £117.81 (33.33% of the amount paid cost).

### 5.1 Amount of work

Figure 5 shows the number of rounds performed by each user, grouped by type of round (*mandatory, extra, extra towards furtherance incentive achievement and extra after furtherance incentive achievement*, as defined in the workflow in Figure 2). The figure does not include a breakdown of mandatory (type $m$) since by design, all 800 users who completed the task submitted either 1 or 11 rounds. Recall that only the last four experimental configurations included furtherance incentives, *i.e.*, have rounds of types $f$ and $x$.

| Round type | *m* | | *e* | | *f* | | *x* | | *all* |
|---|---|---|---|---|---|---|---|---|---|
| Exp. setting | u | r | u | r | u | r | u | r | r |
| 0.1 | 100 | 100 | 99 | 1682 | - | - | - | - | 1782 |
| 0.11 | 100 | 1100 | 99 | 1023 | - | - | - | - | 2123 |
| 1.1 | 100 | 100 | 99 | 1849 | - | - | - | - | 1949 |
| 1.11 | 100 | 1100 | 100 | 1132 | - | - | - | - | 2232 |
| 2.11 MR | 100 | 1100 | 99 | 1011 | 43 | 422 | 17 | 267 | 2800 |
| 2.11 QP | 100 | 1100 | 97 | 1014 | 30 | 211 | 8 | 92 | 2417 |
| 2.11 SB | 100 | 1100 | 96 | 948 | 24 | 192 | 4 | 107 | 2347 |
| 2.11 SP | 100 | 1100 | 96 | 1270 | 14 | 89 | 4 | 40 | 2499 |

Table 4. number of users (u) and overall number of rounds submitted (r) for each experimental setting per round type. Round types are mandatory (m), extra (e), extra to achieve FI reward (f) and extra after achievement of FI reward (x).

The baselines with gamification enabled (1.1 and 1.11) got 10% more extra voluntary rounds than without gamification, however , when looking at the boxplot, the medias are lower, suggesting the difference is due to a handful of outlier users that became very engaged with the game.

Users under experimental setting `2.11 MR` completed the highest median number of rounds (22). On the other hand, users under the two settings requiring only one mandatory round (`0.1` and `1.1`) completed the lowest median number of rounds (12 and 11 respectively). For all other configurations we observe similar median numbers of rounds performed (between 15 and 16). This suggests that money is the most effective furtherance incentive for getting more work done, and that gamification incentives have little to no effect.

Focusing on the extra rounds (type *e*) in the top boxplot of Figure 5, we observe that in all eight experimental settings the vast majority of users completed at least one extra round after the mandatory ones (from 96 out of 100 of `2.11 SB` and `2.11 SP` to the 100 of 1.11). We observe that the two baseline experimental settings with only one mandatory round had a median of extra rounds greater than the baseline settings which required the submission of at least 11 rounds. This suggests that without the implementation of furtherance incentives, users with less mandatory work tend to provide more volunteer work.

We focus now on the rounds performed to achieve a furtherance incentive (type $f$). The first observation concerns the number of users that accepted a furtherance incentive and submitted at least one type $f$ round: e 43 for `2.11 MR` , 30 for `2.11 QP` , 24 for `2.11 SB` and 14 for `2.11 SP` . The median number of rounds performed is less than 11 only for `2.11 SP`(7.5), suggesting that the majority of users who accepted a furtherance incentive offer completed all rounds needed to achieve it.

## 5.2 Working time

Figure 10 shows for each experiment configuration a boxplot of mean time to complete rounds for each user for each type of round. We observe that the two settings requiring only one round have a completion time clearly longer than other settings. It is reasonable to think that this is due to the time taken to familiarise with the interface and the task. Rounds performed after acceptance of the furtherance incentive have lower duration than before acceptance. Users further incentivised by a monetary reward invested less time in those rounds than users further incentivised by non-monetary rewards. On the contrary, users further incentivised by the special power of seeing other user answer took more time to complete the rounds after accepting the incentive. These users also took slightly more time completing the mandatory rounds, which may indicate this users found the task more challenging, that could explain why they chose this particular incentive.

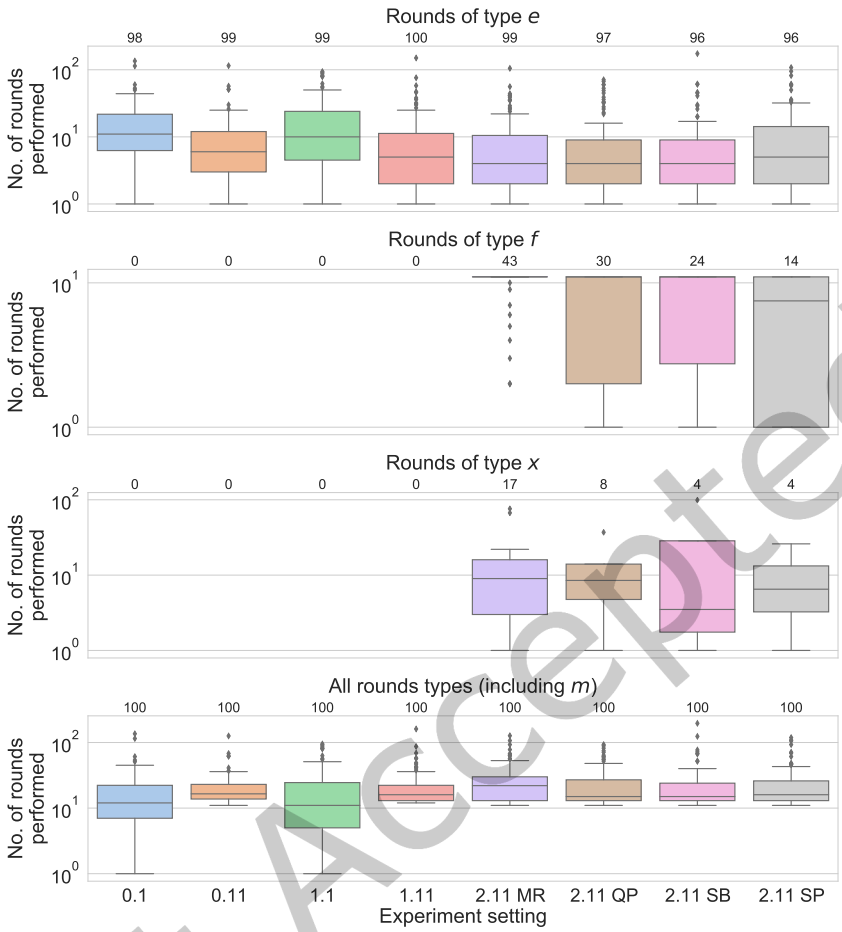



Fig. 5. The number of rounds performed over the eight experiments per round type. Note that we omitted the round type $m$ since by design all 100 users of each experiment settings completed all required rounds, 1 for the 0.1 and 1.1 experiment settings, 11 for the others.



We do note that When comparing tasks with 1 mandatory round versus 11 mandatory rounds, the additional time to recruit workers and for each worker to read the instructions may become significant for large task batches.

## 5.3 Attractiveness of furtherance incentives

Table 5 shows for each furtherance incentive the number of users that were offered the incentive, that accepted it and that completed it. The rightmost three columns show the accepted/offered (A/O) , completed/accepted (C/A) and completed/offered (C/O) ratios. The number of users to which a furtherance incentive was offered is not the same for all incentive types because some users left the task by closing the browser tab instead of using the leave button, preventing the system to offer a furtherance incentive.

We observe that MR is by far the most attractive FI, being completed 35% of the times it was offered, almost double than QP and SB with 18% and 17% respectively and more than four times SPat 8%. This suggests users are

| Incentive | Offered | Accepted | Completed | A/O | C/A | C/O |
|---|---|---|---|---|---|---|
| Money | 96 | 53 | 34 | 0.55 | 0.64 | 0.35 |
| Power | 91 | 33 | 7 | 0.36 | 0.21 | 0.08 |
| Points | 95 | 48 | 17 | 0.51 | 0.35 | 0.18 |
| Badge | 93 | 51 | 16 | 0.55 | 0.31 | 0.17 |

Table 5. The first four columns count the users who: received an offer to perform additional work in exchange for a prize; accepted the offer, and completed it. The last three columns present: the proportion of offers that get accepted, the proportion of accepted offers that get completed, and the overall proportion of users who first accept and then complete previous offers.

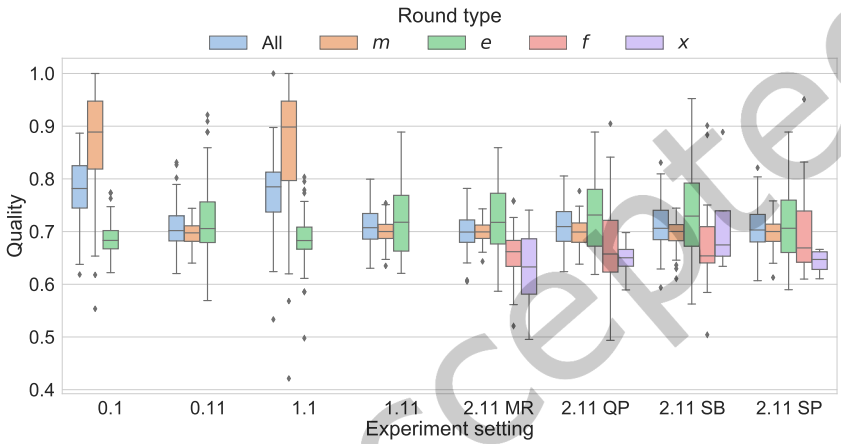



Fig. 6. Quality of user annotations. Each point in the chart is the mean of user quality of each user for rounds of each type for each experimental setting.

more responsive to perform additional work in exchange of a monetary reward. Among the other furtherance incentives, users preferred bonus points QPand a badge SBover the special power SP. This might be due to fact that these two have a lasting impact in score and profile, whereas the special power only lasts until the end of the current task. It is also possible that the points and the badge offer immediate gratification, while the special power does not, it only helps. A third possibility is that the task is easy enough for users to estimate they don't need additional help.

## 5.4 Quality

Figure 6 shows the user annotation quality on the eight different settings and for the four round types. For all configurations and round types, quality oscillates between 0.6 and 0.9. We observe the highest quality for *All* rounds and for mandatory rounds in the settings requiring only one mandatory round (0.1 and 1.1). The quality of the other types of rounds do not show significant differences.

Interestingly, for all four experiment settings involving FIs, the quality of the rounds that come after accepting and completing the FI (namely, types $f$ and $x$) have lower quality than rounds performed before the FI. This suggests that despite FIs motivate users to produce more work, the quality of the additional work is lower. We also observe a reduction in quality after accepting the furtherance incentive, with the largest difference in the monetary reward incentive setting.

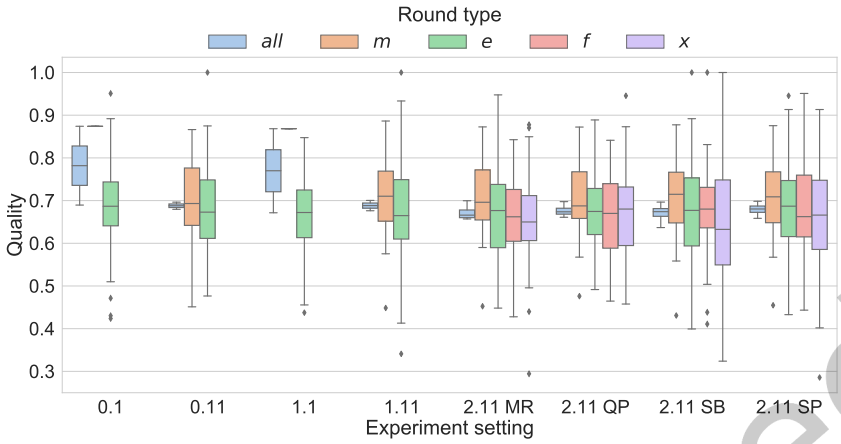



Fig. 7. Quality of the image annotations. Each point in the chart is the mean of the quality of all the annotations collected for an image.

Figure 7 shows the mean of image annotation quality for each image, grouped by round type and experiment setting. Image quality decreases, but less sharply than user quality, indicating that furtherance incentives don't significantly affect the absolute error.

Figure 8 (a) and (b) show the correlation between the number of rounds submitted by each worker and their quality, for experiment settings without FIs and with FIs respectively. In both charts, the highest concentration of points lies between 11 and 25 rounds within 0.6 and 0.7 quality. Quality of short tasks is significantly higher, and furtherance incentives do increase the number of submitted rounds, but with lower quality.

## 5.5 Reliability

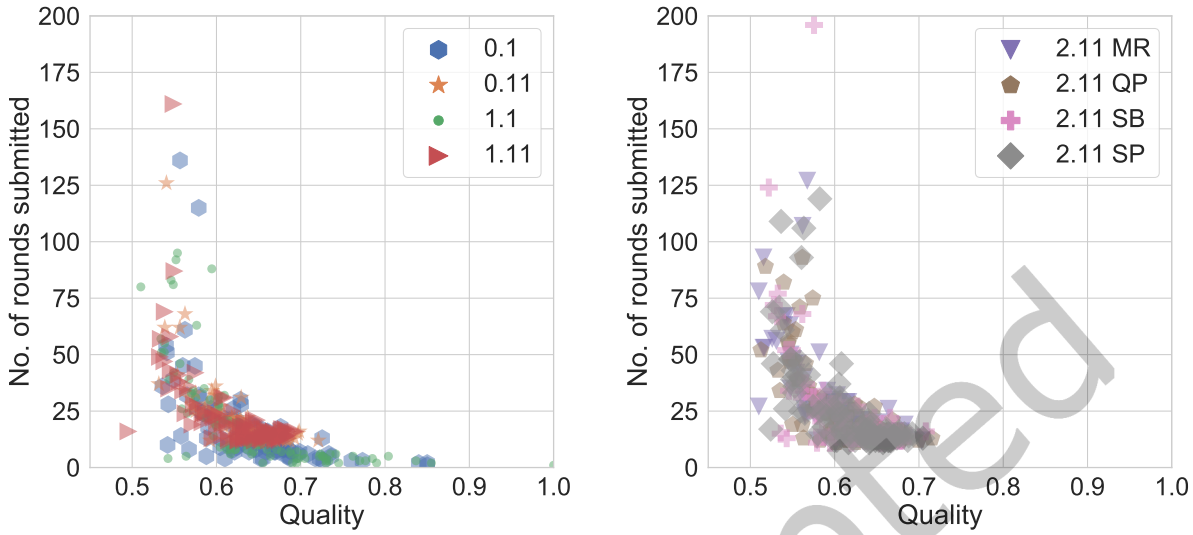
Figure 9 shows the Krippendorff's $\alpha$ values for all round types for all experimental configurations. We could not compute the $\alpha$ on rounds of type $x$ in the `2.11 SP`setting as not enough users submitted additional rounds after completing the furtherance incentive.

We do not observe any significant variation in inter-annotator agreement except for rounds after completing the furtherance incentive (type $x$). This is due to the fact that only 8 users submitted additional rounds of this type for the extra points configuration (`2.11 QP`) and 4 for the special badge configuration (`2.11 SB`).

## 5.6 Cost

Figure 11 (left) compares the mean cost per round with the mean number of rounds for each experimental setting. The monetary incentive leads to more rounds, but at a higher cost. Still, the non-monetary incentives did lead to more rounds than the baselines.

Figure 11 (right) compares the mean cost per round with the mean number of rounds for each experimental setting. This chart highlights the fact that the monetary furtherance incentive leads to a lesser quality than the non-monetary incentives. There is no significant cost/quality difference between the latter.




(a) Experiment settings without FIs

(b) Experiment settings with FIs

Fig. 8. Correlation between number of rounds submitted and the level of quality of each user.

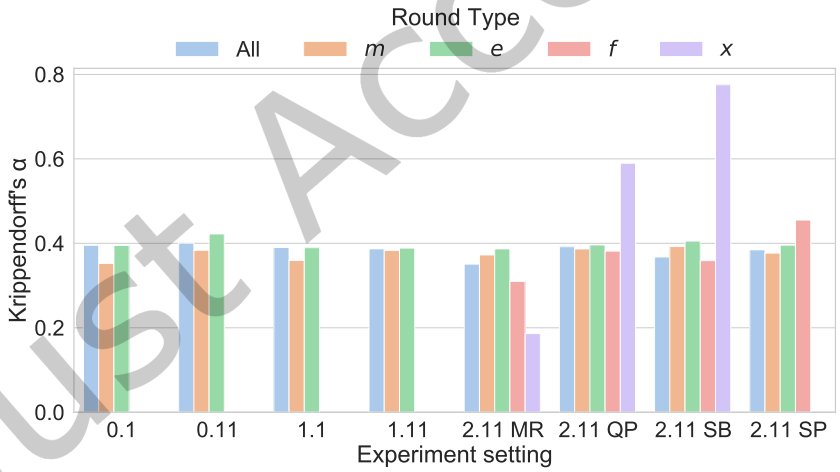



Fig. 9. Krippendorff's $\alpha$ scores for the eight experiment settings. and the four round types. By design, the first four experiment settings do not involve FIs, thus they do not have rounds of type $f$ and $x$.



## 6 DISCUSSION

### 6.1 Furtherance and Participation

Our findings suggest that financial furtherance incentives resulted in the highest number of submissions for each of the furtherance round types, $f$ and $x$. Such financial incentives are commonly employed in crowdsourcing

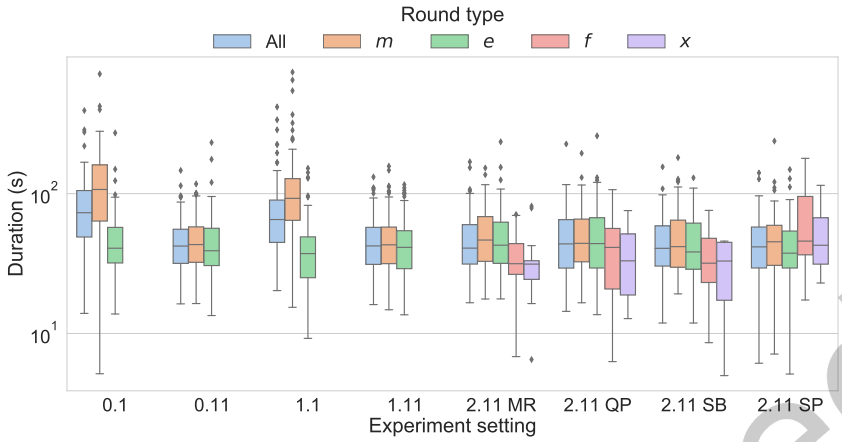



Fig. 10. Each point represents the mean time spent by a user to perform a the rounds of the given type.

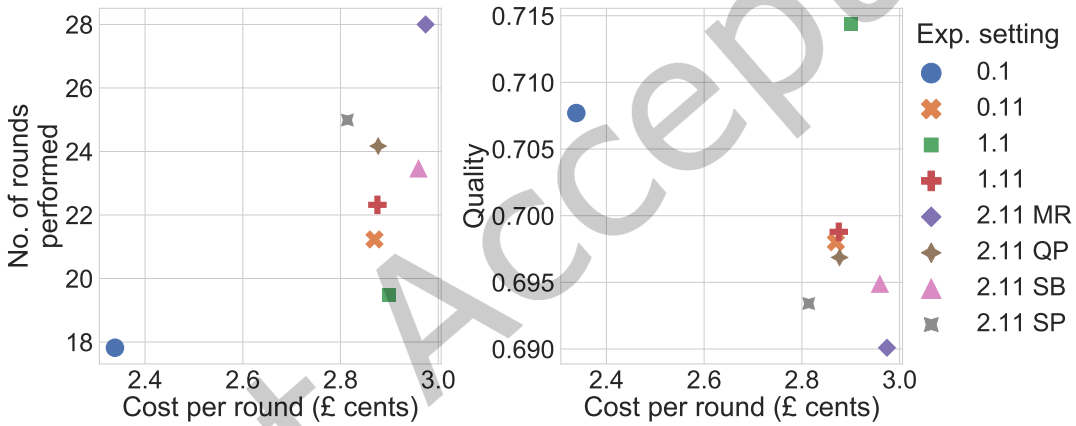



Fig. 11. Association between cost per round and quantity/quality

studies and our results align with previous work in this space (see for example: [3, 13, 53]). It should be noted, however, that for all furtherance conditions, participation was greater in the voluntary *extra* rounds than in the furtherance (*f*) and post-furtherance (*x*) rounds. This similarly echoes previous findings made by [51].

Nevertheless, our findings also partially conflict with prior research in this space. In the domain of crowdsourced games, research has emphasised that both points and badges enhance enjoyment of – and participation within – crowdsourced games [21]. Studies on gamified furtherance incentives have suggested that points, badges and other gamified rewards are far more effective than paid incentives alone [19]. Yet as well as being largely outperformed by financial incentives, our results suggest that users for these conditions were significantly less likely to complete the requisite number of rounds to achieve these rewards.

Our results suggest that popular paid microtask crowdsourcing platform should consider offering gamification as an optional complement to regular working environment. However, further studies should also be conducted to identify relationships between gamification incentives and the cognitive complexity of the task. Our study

was done on a task with a low cognitive complexity, we hypothesise that adding gamification elements to more complex tasks, *e.g.*, the identification of cells with specific features in biomedical images [8], may distract workers and decrease the quality of their contributions. Another interesting further study is to determine if the impact of gamification depends on workers' preferences: some workers may only be interested in monetary reward and may regard gamification elements as superfluous. If this is the case, a practical implementation should also allow workers to enable/disable gamification based on their preferences.

However, it is also important to highlight that unlike the financial and special power conditions, these incentives naturally diminished as workers participated in our experiment. As workers performed more rounds, the number of badges and relative value of points reduced. During the early stages of the task, this is an effective reward structure, relying as it does on relatively rapid gratification [2, 50]. However, as workers gradually contribute more, they reach a stage where subsequent rewards require too high a number of submissions, leading them to potentially abandon the task. In this sense, while these structures may partially drive participant retention and increased engagement, these incentive structures are self-limiting and are ineffective *furtherance* incentives in the mid- to long-term. While this aligns partially with earlier concerns such as those of Kobren et al [28], our findings suggest that these incentives can drive workers to leave tasks even where awards are dynamic.

## 6.2 Data Quality and Reliability

While the furtherance incentives used throughout our experiments were generally at least partially effective at increasing participation and increasing the *quantity* of submissions made by workers, this increased participation came at the cost of the *quality* of worker submissions. Indeed, workers who submitted the greatest numbers of submissions also achieved the lowest levels of submission quality.

This was particularly true of financial incentives, which were associated with both lower levels of quality overall and the most significant association between increased quantity and reduced quality. In itself, this conclusion is unsurprising and aligns with previous studies which have explored the impact of financial incentives on the results of crowdsourcing studies [33, 43]. Nevertheless, there is some evidence within the wider literature that financial furtherance incentives can lead to improved quality where these incentives are offered conditionally, dynamically and *crucially* when they are paired with the need to submit high quality submissions [57].

Moreover, this reduction in data quality extends to inter-rater agreement. All three gamified experiment types – points, badges and functionality – achieved similar levels of inter-rater reliability, with furtherance incentives resulting in higher agreement in the case of the special power condition and only a minor reduction in agreement for points and badges. For financial incentives, however, monetary rewards were associated with significantly lower levels of inter-rater reliability and this was particularly the case once the furtherance incentive was introduced. Prior to this point, agreement was comparable with that seen in other experimental conditions. This suggests that workers that accepted non-monetary incentives agree more, leading to higher quality collected data. We also note that the agreement for extra voluntary rounds (type $x$) for non-monetary incentives is higher than for monetary rewards. This suggests that workers that took the money lose focus, maybe in an attempt to maximise profit, while workers that keep working for gamification incentives remain engaged with the task.

For task designers, this implies that monetary rewards is a powerful furtherance incentive, but the agreement decrease means that more complex aggregation functions may be required to ensure good quality data.

Additionally, our experiment focused largely on extrinsic motivations, either in the form of financial incentives or gamification rewards such as points, badges and functionality. Even so, it is important to highlight the importance of *intrinsic* motivations in encouraging participation in crowdsourcing activities, particularly as a furtherance incentive. Workers' natural curiosity about tasks and their outcomes has been demonstrated to be an effective furtherance incentive and which notably – and unlike many of our incentives – is not associated with a

reduction in quality [31]. Indeed, an increase in extrinsic rewards and particularly increased financial incentives may limit the drive to perform tasks well as associated with *intrinsic* motivations [25].

### 6.3 Implications of Furtherance Incentives

One interesting observation within our findings is the significant decrease in annotation quality associated with extra rounds in comparison to mandatory rounds. Introducing furtherance incentives was associated with lower annotation quality both for individual workers and for images overall, even where prior to the introduction of these incentives average user annotation quality actually increased. This suggests that as workers are offered incentives to maximise their contributions, there is a resulting trade-off between quantity and quality, leading workers to focus solely on making as many submissions as they can to earn further rewards. Indeed, average completion time similarly fell after incentives were offered, suggesting a shift in worker behaviour. Regardless of reasoning, our results suggest a strong association between the number of rounds a user submits and their overall quality.

This is a key area for further study, which should consider whether it is simply the provision of incentives themselves or perhaps an associated confounding factor – for example, time spent on a task – that leads to this fall in quality. We note that while prior research has identified a fatigue effect associated with increasing engagement with crowdsourcing tasks, the negative impacts of this effect are generally offset by increased task familiarity [58]. Nevertheless, on this basis, we recommend that furtherance incentives are poorly suited to quality-critical tasks and such contexts would be better suited by maximising overall worker numbers rather than the contributions made by each individual.

Additionally, while the financial incentive was most effective at driving increased participation, we note that the cost-effectiveness of this participation is an area of concern. For the cost of a user performing 11 rounds, workers who agreed to the furtherance incentive completed on average just 16 rounds per worker and this number was significantly inflated by a small number of highly active workers. This was significantly less than the average of 21 rounds completed by workers in this condition prior to the introduction of the monetary incentive.

On this basis, we can assume that a greater number of rounds would be achieved by simply recruiting new workers, rather than offering furtherance incentives to existing workers.

Therefore, while gamified incentives may appeal to only a subset of workers and are less popular than financial incentives, they are more cost-effective and show a smaller quality degradation. A requester who would like to implement furtherance incentives should take into account the balance between a higher cost but more quantity of work at a lower quality and inter-rater agreement provided by monetary incentives, and a lower cost, higher quality but less quantity of work associated to gamified incentives.

### 6.4 Generalisability

In terms of generalisability to other types of tasks, our study did not test association between task features and the performance of furtherance incentives. We hypothesise that there is a difference between tasks that expect objective answers (counting, True/False) and tasks that expect subjective answers (sentiment judgement). For the latter, inter-annotator agreement cannot be used to estimate worker quality in real time due to the fact that the task does not expect full agreement. We also believe that for tasks that offer workers intrinsic benefits (altruism, sense of belonging, educational) all gamified configurations will be effective, while on the contrary repetitive or uninteresting tasks only the extrinsic incentive configuration (monetary reward).

We also think that worker profiles impact the effectiveness of Qrowdsmith approach in the same way they affect other types of Crowdsourcing platforms [14, 42]. Experiments to measure this impact are interesting future work.

Finally, we also hypothesise that task duration impacts the effect of gamified incentives. If a worker waits too long to see their points increasing or getting badges, the power of the incentive is diminished. For those cases, we believe it would be beneficial – where possible – to subdivide larger tasks into smaller units to increase the reward frequency.

### 6.5 Limitations

We note a number of limitations associated with our study, due in part to the use of a live crowdsourcing platform. Firstly, we chose to allow workers to complete as many voluntary rounds ($e$) as they wished, to allow us to compare the impact of incentives with worker engagement prior to their introduction. However, this meant that it was not possible to compare for the number of rounds workers had completed before the furtherance incentive was introduced, which may potentially have led to workers suffering different levels of fatigue across and within conditions. While this is unlikely to have significantly impacted our findings – workers from conditions with the lowest number of extra rounds did not perform significantly more work – it is nonetheless an important consideration in further work and analysis.

An additional limitation was the nature of payments within the Prolific platform. While Prolific asks requesters to price their tasks based on completion time, workers are generally paid a flat rate for performing a given task. As a result, as workers spent more time on our task, their effective level of pay decreased. On the one hand, this was again unlikely to influence our study given that workers showed a great willingness to continue to contribute without any further pay being offered. Nevertheless, it is a potential confounding factor that should be explored in further work.

Many of Qrowdsmith's gamification features were reliant on the engagement of a large number of participants. Since we were keen not to introduce artificial features, engagement with the chat, leaderboard and alert features was all highly dependent on the number of workers contributing at a given time. We were unable, therefore, to control for these levels of participation and cannot rule out the possibility that individual volunteers' decisions to contribute or leave the task were influenced by the number of participating individuals at any given time.

## 7 CONCLUSIONS AND FUTURE WORK

In this paper, we presented Qrowdsmith, a platform for gamifying microtask crowdsourcing that enables the use of furtherance incentives in crowdsourcing tasks. We used Qrowdsmith to analyse the impact of gamified furtherance incentives on the quantity and quality of contributions to an image labelling task. We studied eight experimental conditions, varying the number of rounds workers were expected to perform, the inclusion or exclusion of gamification elements, the payment offered and the inclusion of furtherance incentives in the form of payment, points, badges or game features.

Our results suggest that monetary rewards were the most effective at encouraging worker retention and increasing engagement over gamified incentives such as points and badges. Nevertheless, these financial incentives were also associated with the lowest submission quality. Additionally, all furtherance incentives were associated with lower numbers of voluntary additional submissions, but with a higher submission quality.

Overall, our results suggest that while gamified elements can play a role in encouraging users to contribute further to tasks, their effect is limited compared to additional payments. Finally, we note the implication from our work that requesters may be better served using their limited resources to recruit additional workers, rather than to attempt to encourage existing workers to contribute for longer.

Further work could also compare different reward approaches. The ones we used based on the quantity of work submitted against strategies based on features like annotation quality, time spent per task, or the agreement with previous annotations. Another interesting direction is the study if our results generalise to different types of tasks beyond image annotation.

**Acknowledgements:** This work was partially supported by the European Union's Horizon 2020 research and innovation programmes Qrowd and Action, under grant agreements No 732194 and No 824603; and Cleopatra, under the Marie Skłodowska-Curie grant agreement No 812997.

## REFERENCES

[1] Gregory Afentoulidis, Zoltán Szlávik, Jie Yang, and Alessandro Bozzon. 2018. Social gamification in enterprise crowdsourcing. In *Proceedings of the 10th ACM Conference on Web Science*. 135–144.

[2] Adam Atkins, Vanissa Wanick, and Gary Wills. 2017. Metrics Feedback Cycle: measuring and improving user engagement in gamified eLearning systems. *International Journal of Serious Games* 4, 4 (2017). https://doi.org/10.17083/ijsg.v4i4.192

[3] Elena M Auer, Tara S Behrend, Andrew B Collmus, Richard N Landers, and Ahleah F Miles. 2021. Pay for performance, satisfaction and retention in longitudinal crowdsourced research. *Plos one* 16, 1 (2021), e0245460.

[4] Shahzad Sarwar Bhatti, Xiaofeng Gao, and Guihai Chen. 2020. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey. *Journal of Systems and Software* 167 (2020), 110611.

[5] Alice M Brawley and Cynthia LS Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531–546.

[6] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In *HCOMP*.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Vincenzo Della Mea, Eddy Maddalena, Stefano Mizzaro, Piernicola Machin, and Carlo A. Beltrami. 2014. Preliminary results from a crowdsourcing experiment in immunohistochemistry. *Diagnostic Pathology* 9, 1 (2014), S6. https://doi.org/10.1186/1746-1596-9-S1-S6

[9] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. Association for Computing Machinery, New York, NY, USA, 469–478. https://doi.org/10.1145/2187836.2187900

[10] Xuefei Nancy Deng and KD Joshi. 2016. Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions. *Journal of the Association for Information Systems* 17, 10 (2016), 3.

[11] Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying crowd workers for collaborative work. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[12] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[13] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[14] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-Crowd: Tell Me What You like, and i'll Tell You What to Do. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 367–374. https://doi.org/10.1145/2488388.2488421

[15] Alexandra Eveleigh, Charlene Jennett, Stuart Lynn, and Anna L Cox. 2013. "I want to be a captain! I want to be a captain!" gamification in the old weather citizen science project. In *Proceedings of the first international conference on gameful design, research, and applications*. 79–82.

[16] Yuanyue Feng, Hua Jonathan Ye, Ying Yu, Congcong Yang, and Tingru Cui. 2018. Gamification artifacts and crowdsourcing participation: Examining the mediating role of intrinsic motivations. *Computers in Human Behavior* 81 (2018), 124–136.

[17] Oluwaseyi Feyisetan and Elena Simperl. 2017. Social incentives in paid collaborative crowdsourcing. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 6 (2017), 1–31.

[18] Oluwaseyi Feyisetan and Elena Simperl. 2019. Beyond Monetary Incentives: Experiments in Paid Microtask Contests. *ACM Transactions on Social Computing* 2, 2 (March 2019), 6. https://doi.org/10.1145/3321700

[19] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th International Conference on World Wide Web*. 333–343.

[20] Karan Goel, Shreya Rajpal, and Mausam Mausam. 2017. Octopus: A framework for cost-quality-time optimization in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[21] Dion Hoe-Lian Goh, Ei Pa Pa Pe-Than, and Chei Sian Lee. 2017. Perceptions of virtual reward systems in crowdsourcing games. *Computers in Human Behavior* 70 (2017), 365–374.

[22] Zhuojun Gu, Ravi Bapna, Jason Chan, and Alok Gupta. 2021. Measuring the Impact of Crowdsourcing Features on Mobile App User Engagement and Retention: A Randomized Field Experiment. *Management Science* (2021).

[23] L Han, K Roitero, U Gadiraju, C Sarasua, A Checco, E Maddalena, and G Demartini. 2018. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM.
[24] Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S Bernstein. 2017. A glimpse far into the future: Understanding long-term crowd worker accuracy. *CSCW: Computer-Supported Cooperative Work and Social Computing* (2017).
[25] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*. 143–154.
[26] Aikaterini Katmada, Anna Satsiou, and Ioannis Kompatsiaris. 2016. Incentive mechanisms for crowdsourcing platforms. In *International conference on internet science*. Springer, 3–18.
[27] Melissa G Keith, Peter Harms, and Louis Tay. 2019. Mechanical Turk and the gig economy: exploring differences between gig workers. *Journal of Managerial Psychology* (2019).
[28] Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th international conference on world wide web*. 592–602.
[29] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).
[30] Jeffrey Laut, Francesco Cappa, Oded Nov, and Maurizio Porfiri. 2017. Increasing citizen science contribution using a virtual peer. *Journal of the Association for Information Science and Technology* 68, 3 (2017), 583–593.
[31] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A Terry, and Krzysztof Z Gajos. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4098–4110.
[32] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2017. Measuring the effect of" bottom-up" gamification in a microtask setting. In *Proceedings of the 21st International Academic Mindtrek Conference*. 63–72.
[33] Jae-Eun Lim, Joonhwan Lee, and Dongwhan Kim. 2021. The effects of feedback and goal on the quality of crowdsourcing tasks. *International Journal of Human–Computer Interaction* (2021), 1–13.
[34] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. 2015. The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior research methods* 47, 2 (2015), 519–528.
[35] Eddy Maddalena, Luis-Daniel Ibáñez, and Elena Simperl. 2020. Mapping Points of Interest Through Street View Imagery and Paid Crowdsourcing. *ACM Trans. Intell. Syst. Technol.* 11, 5, Article 63 (Aug. 2020), 28 pages. https://doi.org/10.1145/3403931
[36] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/3121050.3121060
[37] David Martin, Sheelagh Carpendale, Neha Gupta, Tobias Hoßfeld, Babak Naderi, Judith Redi, Ernestasia Siahaan, and Ina Wechsung. 2017. Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. In *Evaluation in the crowd. crowdsourcing and human-centered experiments*. Springer, 27–69.
[38] Winter Mason and Duncan J Watts. 2009. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*. 77–85.
[39] Benedikt Morschheuser, Juho Hamari, Jonna Koivisto, and Alexander Maedche. 2017. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies* 106 (2017), 26–43.
[40] Benedikt Morschheuser, Juho Hamari, and Alexander Maedche. 2019. Cooperation or competition–When do people contribute more? A field experiment on gamification of crowdsourcing. *International Journal of Human-Computer Studies* 127 (2019), 7–24.
[41] Sara Moussawi and Marios Koufaris. 2015. Working on low-paid micro-task crowdsourcing platforms: An existence, relatedness and growth view. (2015).
[42] Jabu Mtsweni, Ernest Ketcha Ngassam, and Legand Burge. 2016. A profile-aware microtasking approach for improving task assignment in crowdsourcing services. In *2016 IST-Africa Week Conference*. 1–10. https://doi.org/10.1109/ISTAFRICA.2016.7530702
[43] Fábio R Assis Neto and Celso AS Santos. 2018. Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management. *Information Processing & Management* 54, 4 (2018), 490–506.
[44] Mads Kock Pedersen, Nanna Ravn Rasmussen, Jacob Sherson, and Rajiv Vaid Basaiawmoit. 2017. Leaderboard Effects on Player Performance in a Citizen Science Game. In *European Conference on Games Based Learning*. Academic Conferences International Limited, 531–537.
[45] Julien Pilourdault, Sihem Amer-Yahia, Dongwon Lee, and Senjuti Basu Roy. 2017. Motivation-Aware task assignment in crowdsourcing. In *20th International Conference on Extending Database Technology, EDBT 2017*. OpenProceedings. org, 246–257.
[46] Sihang Qiu, Alessandro Bozzon, Max V Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.
[47] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[48] Neal Reeves, Peter West, and Elena Simperl. 2018. A game without competition is hardly a game": The impact of competitions on player activity in a human computation game. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[49] Elena Simperl, Neal Reeves, Chris Phethean, Todd Lynes, and Ramine Tinati. 2018. Is virtual citizen science a game? *ACM Transactions on Social Computing* 1, 2 (2018), 1–39.
[50] David Smahel, Lukas Blinka, and Ondrej Ledabyl. 2008. Playing MMORPGs: Connections between Addiction and Identifying with a Character. *CyberPsychology & Behavior* 11, 6 (2008), 715–718. https://doi.org/10.1089/cpb.2007.0210 PMID: 18954271.
[51] Sofia Eleni Spatharioti, Rebecca Govoni, Jennifer S Carrera, Sara Ann Wylie, and Seth Cooper. 2017. A Required Work Payment Scheme for Crowdsourced Disaster Response: Worker Performance and Motivations.. In *ISCRAM*.
[52] Kai Spindeldreher and Daniel Schlagwein. 2016. What Drives the Crowd? A Meta-Analysis of the Motivation of Participants in Crowdsourcing.. In *PACIS*. 119.
[53] Elizabeth Stoycheff. 2016. Please participate in Part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations* 9 (2016), 2059799116672879.
[54] Fei-Yue Wang, Kathleen M. Carley, Daniel Zeng, and Wenji Mao. 2007. Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligent Systems* 22, 2 (2007), 79–83. https://doi.org/10.1109/MIS.2007.41
[55] Congcong Yang, Yuanyue Feng, Xizhi Zheng, Ye Feng, Ying Yu, Ben Niu, and Pianpian Yang. 2018. Fair or not: Effects of gamification elements on crowdsourcing participation. In *Proceeedings of 18th International Conference on Electronic Business*. 325–335.
[56] Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
[57] Ming Yin and Yiling Chen. 2015. Bonus or not? learn to reward in crowdsourcing. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
[58] Ying Zhang, Xianghua Ding, and Ning Gu. 2018. Understanding fatigue and its impact in crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 57–62.
[59] 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.