Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives

Oluwaseyi Feyisetan
University of Southampton
Southampton, UK
oof1v13@soton.ac.uk

Elena Simperl
University of Southampton
Southampton, UK
e.simperl@soton.ac.uk

Max Van Kleek
University of Southampton
Southampton, UK
emax@ecs.soton.ac.uk

Nigel Shadbolt
University of Southampton
Southampton, UK
nrs@ecs.soton.ac.uk

ABSTRACT

Crowdsourcing via paid microtasks has been successfully applied in a plethora of domains and tasks. Previous efforts for making such crowdsourcing more effective have considered aspects as diverse as task and workflow design, spam detection, quality control, and pricing models. Our work expands upon such efforts by examining the potential of adding gamification to microtask interfaces as a means of improving both worker engagement and effectiveness. We run a series of experiments in image labeling, one of the most common use cases for microtask crowdsourcing, and analyse worker behavior in terms of number of images completed, quality of annotations compared against a gold standard, and response to financial and game-specific rewards. Each experiment studies these parameters in two settings: one based on a state-of-the-art, non-gamified task on CrowdFlower and another one using an alternative interface incorporating several game elements. Our findings show that gamification leads to better accuracy and lower costs than conventional approaches that use only monetary incentives. In addition, it seems to make paid microtask work more rewarding and engaging, especially when sociality features are introduced. Following these initial insights, we define a predictive model for estimating the most appropriate incentives for individual workers, based on their previous contributions. This allows us to build a personalised game experience, with gains seen on the volume and quality of work completed.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

algorithms; experimentation; theory

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741639.

Keywords

crowdsourcing; microtasks; gamification; incentives engineering

1. INTRODUCTION

Crowdsourcing offers a powerful and scalable way of outsourcing knowledge-intensive tasks to a large group of remote contributors over the Internet. Eight years since the concept was first introduced [10], it seems to have found global adoption; virtually any industry, science discipline, or public sector agency could tell a story about how they reached out to the wisdom of the crowds to improve their services and react more flexibly to customer demand, run comprehensive data collection and analysis projects, or collect ideas and views for a better informed policy making [3]. However, with every such success story, it has also become apparent that running crowdsourcing projects effectively is not straightforward and requires in-depth insight into a wide range of topics, including interface design, statistical data analysis, and incentives engineering [15, 16].

One of the most popular forms of online crowdsourcing are paid microtasks. People have tried to understand how to make microtask projects more effective, looking at aspects as diverse as task and workflow design, spam detection, quality control, and pricing models. Our work shares similar aims. We explore the use of gamification in combination with paid microtasks as a means to improve both task performance and worker experience.

We run a series of experiments in image labeling, one of the most common use cases for paid microtask crowdsourcing[1] on platforms such as Amazon Mechanical Turk[2] and CrowdFlower[3], and analyse worker behavior in terms of number of images completed, quality of annotations compared against a golden standard, as well as monetary and game-specific rewards. Each experiment studies these parameters in two settings: one based on a state-of-the-art, non-gamified 'job' on CrowdFlower (i.e., the unit of work on this platform); and another one using an alternative interface incorporating several game elements. The second setting uses CrowdFlower as well, but only to seek contributors; it offers the same reward for the same amount of work as the baseline task, but points to an external page where the gamified version of the task is deployed. More specifically, in the second condition, CrowdFlower

[1]See `http://www.behind-the-enemy-lines.com/2010/10/what-tasks-are-posted-on-mechanical.html`.
[2]Amazon Mechanical Turk - `https://www.mturk.com`
[3]CrowdFlower - `http://www.crowdflower.com`

workers are asked to engage with a game-with-a-purpose (GWAP) [27] called Wordsmith, which was developed for the purpose of these experiments.[4]

When creating Wordsmith, our primary focus was not on coming up with a fundamentally novel game experience to collect image labels, but on building an experimental framework to test our research hypotheses regarding the interplay between monetary and gamification-centric rewards. Overall, our design choices were guided by the four traits that define games: a goal, rules, a feedback system, and voluntary participation [21]. The game elements we implemented are some of the most common in the literature [31]: points, levels, leaderboards, and badges. These components provide feedback and reinforcement, while helping workers judge their progress relative to the end-goal of the task, as well as relative to other workers. We also included progress bars and alerts, which informed on their proximity to level goals. Finally, we put in place elements that enforced the game rules, for example, modules that check for spam, duplicates, and non-English words.

Our basic hypothesis is that by designing a playful interface for the image labeling task - as opposed to the functional style common to most microtask platforms - we will encourage workers to engage with the task more, independently of the actual monetary reward. This hypothesis was confirmed by our findings, which revealed better accuracy (an improvement of almost 10% compared to the baseline condition) and significantly lower costs per annotated image (5,708 unique labels collected via the game vs. 111 unique labels contributed through equivalent, non-gamified microtasks, see Experiment 1 in Section 5). We tested this hypothesis on different variations of image labeling tasks, in which we increased the complexity of the task and adjusted the prices accordingly, observing a similar trend.

Then we looked into the impact of different game elements and related incentives on the behavior of the workers, following the *SAPS framework* (Status, Access, Power, Stuff) presented by [31]. Besides studying how people responded to the primary gamification components (leaderboard, levels, points, and badges), we also introduced a sociality aspect, which was originally missing from the game and allowed contributors to view each others' achievements. Our final incentive dimension was additional cash payment for more work, once the goals of the initial task had been achieved.

These new experiments produced clear evidence of the positive effects of game mechanics on both task performance and crowd engagement; up to five times more unique labels were generated while preserving a comparable level of accuracy (see Experiment 2 in Section 5). This is true particularly for sociality features, which are largely absent on microtask platforms.

Following these insights, we went on to create a predictive model that estimates the most suitable set of incentives for individual workers, based on their previous contributions. This allowed us to build a personalized game experience, with positive results on the volume and quality of work completed. With this model, we were able to obtain 19% more concordant image keywords (4849 vs. 4091) while maintaining the same average pair-wise agreement score. We also recorded a significant uptake in image tagging (with 77% of players tagging an extra image when confronted with targeted incentives against 27% in the experiments using a randomly selected incentive).

Overall, the results of these experiments shed light on possible design improvements of paid microtasks environments in order to achieve better task performance and make the overall experience more fair and rewarding for the workers. While we are not necessarily arguing for a fully-fledged gamification of microtask platforms, considering specific game mechanics [31], or in fact, any social design features that are widely discussed on online communities literature [17] is worth further investigations. This is important not just for purely utilitarian motives on the side of the task requesters, but also in the context of the ongoing debate on ethical and fair crowdsourcing [11].

Previous work has approached such aspects through studies of crowd motivations [14], discussing the rich repertoire of extrinsic and intrinsic reasons that drive people to contribute to microtask projects. Our experiments quantify some of these insights. We deliberately chose a task that is well-known in the crowdsourcing literature, as we were aiming for task-independent findings, which were only minimally influenced by interface or quality control aspects. For the same reasons, we opted for average market prices to reward participation; lower pays would have been less attractive (and unfair) for workers, higher ones might have appealed to people who were primarily financially incentivised. We believe more research needs to be done to build microtask platforms that reflect and support the values and motivations of the crowd as an integral part of their functionality. Our experiments give evidence that such efforts could be be beneficial both workers, and for requesters.

2. BACKGROUND

In this section, we briefly review some of the most relevant prior work pertaining to maximising the effectiveness of incentivised crowdsourcing. In particular, we focus on approaches that use game mechanics in human computation, and on methods that aim to optimize the performance of crowd workers, be that by offering bespoke incentives, or by assigning tasks to those workers who are likely to be able or willing to complete them accurately. As much of this background literature is inspired by, and explained using, theories of human motivation, we touch on fundamental work in that space as well.

2.1 Motivation and incentives

While efforts at designing successful crowdsourcing projects have considered a variety of dimensions, including end-user interfaces, spam detection, and quality control, some of the most influential works in recent crowdsourcing literature have approached this problem by looking at crowd engagement. This is seen as an effective way to achieve better productivity and ensure the sustainability of crowdsourcing platforms over time. Research on crowd engagement covers various aspects, from studies of motivations of contributors to specific projects to applications of theoretical models from economics to the newer scenarios of online labor markets.

Intrinsic and extrinsic motivation Frameworks such as the Self Determination Theory [4] make the distinction between *intrinsic* and *extrinsic* motivation to study the reasons why individuals decide to contribute to a task. The former covers those types of tasks which are perceived as interesting, enjoyable, or rewarding in themselves, and is seen as responsible for engagement in many on and off-line activities, including reading books, socialisation, or participation in virtual worlds and volunteering projects. Extrinsic motives, by contrast, are related to factors that are not inherently related to the actual task, but are appealing for some external reason such as status or reputation within a group. Equally powerful, they underpin a fair share of our contributions to social networks, discussion forums, or community projects.

Theories of external reward and incentivisation
Yet for many classes of tasks, the most reliable way for getting crowd workers engaged remains through explicit external rewards, in particular monetary payment [13]. However, studies have shown

[4]Wordsmith - `http://seyi.feyisetan.com/wordsmith`

that such models are no guarantee for effective crowdsourcing. Where in some cases higher rewards are likely to expedite task completion, they do not necessarily lead to an improved worker performance and attract more spammers [20, 19, 30]. Such scenarios have to consider anchoring effects, caused by discrepancies between perceived work value and actual remuneration, and drop-off effects, where they stop after hitting their self-defined targets no matter how high the additional gains might be [20, 30].

Gamification and its effects Gamification refers to "the use of game design elements in non-game contexts" [7]. It often includes adding game-like rewards, and may also include competitive and social elements, such as leaderboards, explicit competitions, as well as group and individual performance feedback to encourage engagement. Many projects have already demonstrated substantial success in applying this idea to crowdsourcing settings; this applies most prominently for *games-with-a purpose* (GWAPs) [27], which build a game narrative around human computation tasks such as image labeling [26], protein folding,[5] or language translation.[6]

Similarly to the concerns raised in the context of external rewards and incentivisation [18], gamification has been seen, in some context, to undermine intrinsic benefits by subjugating and trivialising contributions into simple game goals and achievements.[7] This effect has been called *overjustification* and has been the subject of various studies with intriguing results; while some negative effects of overjustification have been recurrently reproduced, current research acknowledges the fact that its prevalence seems to be highly dependent on context and that, in most cases, extrinsic rewards complement rather than hamper intrinsic motivations for participating [5, 22].

Comparative studies
Several works have explored such effects by undertaking comparative studies of different forms of crowdsourcing to solve specific tasks. In [25] the authors analysed paid microtasks and a game-with-a-purpose to build a conceptual schema from topics of Wikipedia articles. They reported similar levels of task performance for both approaches, and were concerned by the costs associated with the development and maintenance of a sustainable community of GWAP players. Similarly, a comparison of a purely gamified environment (with no monetary reward) vs standard paid microtasks by Jurgens et al. showed that while comparable quality was achieved from both conditions, the former yielded an overall cost savings (when game development costs were factored in) of 63% overall to achieve task completion [12]. However, the cost savings came at an expense of timely completion; results from the paid microtask platforms were completed in hours, while results from the game approach (Puzzle Racer) trickled in over two weeks.

Meanwhile, a few previous studies have also looked at supplementing monetary reward with gamification, as we do in this paper. A study of crowdsourced judgements on relevance assessments and clustering compared traditional HITs and gamified HITs across the dimensions of quality (compared against a gold standard), efficiency (time required to collect judgements) and incentives (financial vs fun)[8]. Their results show that adding gamification yielded quicker judgements at a higher quality over purely financial incentives. Our work offers a more in-depth study of different types of incentives, alongside a model to predict the most effective type of reward. Other relevant studies in literature include a comparison of crowd-based, game-based, and machine-based approaches by [9].

[5]Fold It - `http://fold.it`
[6]Duolingo - `https://www.duolingo.com/`
[7]Criticisms of Gamification - `http://radar.oreilly.com/gamification-criticism`

2.2 Methods to make crowdsourcing more effective

While in the previous section we primarily looked at prior studies of motivation and incentives aspects in particular scenarios, we will now give an overview of methods which help make crowdsourcing projects more effective by optimizing key components of such projects.

A number of descriptive frameworks have been proposed in the literature to capture the nuances of incentives engineering beyond simplistic *'fun or money'* considerations. Some of these include *MICE* (Money, Ideology, Coercion, Excitement) [2], *RASCLS* (Reciprocation, Authority, Scarcity, Commitment, Liking, Social Proof) [2], and *SAPS* [31]. SAPS represents Status, Access, Power and Stuff, intended to represent a system of incentives from the most desired to the least desired, and the cheapest to the most expensive. We adopted this framework in our experiments.

Mechanisms for effective allocation of incentives have been studied in market and auction platforms, wireless and peer-to-peer networks and corporate organisations. In the context of crowdsourcing, a number of studies have been carried out, applying *game-theory* techniques to incentive design [28, 29]. These two pieces of work focus on financial incentives and a premise of inter-player strategy dependency. Not all crowdsourcing tasks can be modelled in this way; we adopt a probabilistic approach based on prior player behaviours to predict appropriate incentives beyond the purely financial. Similar techniques are used for various purposes in crowdsourcing design, in particular to inform the assignment of tasks to workers or to predict task completion [6, 24].

A large body of work has been dedicated to task and workflow design, as well as quality control (see, for instance, [23] for a recent compilation). We take their findings into account when implementing the basic interfaces published on CrowdFlower as well as the means to check quality and validate results.

3. CROWDSOURCING IMAGE LABELING: OUR APPROACH

In this section, we introduce a high level overview of our approach to crowdsource image labeling. We present our microtask design model and strategies for undertaking crowd work. This involves the use of an external platform, CrowdFlower, and our bespoke game *Wordsmith*.

3.1 Task model

We now describe our model for maximising the output from crowd assigned tasks while maintaining quality.

Task. Each HIT (Human Intelligence Task) consists of n images, $x = \{x_1, ..., x_n\}$, which can each be described by a set of m keywords $k = \{k_1, ..., k_m\}$, where m is a large unknown number. Each task seeks to capture new keywords that correctly describes each image.

Requester. The requester desires to have as many image annotations as possible, without compromising on the quality of the describing keywords. The requester requires the help of human agents to carry out the tasks.

Strategy. We define two requester strategies for presenting tasks. The *crowd* strategy relies on traditional crowdsourcing techniques in a standard "image field - text fields" layout. The *game* strategy employs game mechanisms, and a game based interface to capture keywords. Our nomenclature defines human agents in the crowd strategy as *workers*, and those in the game as *players*.

Crowd → Worker. Each worker provides judgement on a task by assigning m keywords $\{k_1, ..., k_m\}$ to n images $\{x_1, ..., x_n\}$ in

a traditional crowdsourcing system. We used CrowdFlower as our crowdsourcing platform, presenting each task using the standard image annotation template provided. In this strategy, n and m are defined and fixed by the requester.

Game → Player. Each player provides judgements on a task in a fashion similar to workers in the *crowd* strategy. However, in the *game* strategy, players can tag a variable number of images as they progress through more levels.

Quality. Is defined by consensus. The number of keywords matching a quasi-gold standard bank, gives an overview of the quality of annotations. This was extended to also cover consensual annotations within workers and players - as this suggests probable new keywords for the image.

3.2 Recruitment

We sourced all our human agents from CrowdFlower. For each experiment, we created 2 jobs which channelled task resources to the crowd strategy and game strategy. We used identical settings for each experiment set, consisting of the following parameters:

Geography - limited to the top 15 English speaking countries, and the top CrowdFlower contributor countries.
Skills - we chose *Level 2 Contributors*, which account for 36% of monthly judgements.
Judgements - 3 per unit, which meant each image would be annotated by at least 3 human agents.
Behaviour - each human agent was paid for 1 task, i.e., paid to tag m images, with n keywords, each as determined by the requester. For this, CrowdFlower tracks the IPs and aliases created by the agents.
Reward/Time Limits - reward payment and completion time limits were experimentally set as described later.

3.3 Game design

We designed an image-labelling game called *Wordsmith* for our experimental game conditions. The design was heavily borrowed from the ESP game, with variations described below. The basic elements of Wordsmith consisted of an image frame and text fields for inputting keywords. We describe Wordsmith in terms of the four defining properties of games [21]) as follows: its goal, rules, feedback mechanisms and participation.

Game goal
The goal of the game was to annotate as many images as possible (up to the maximum in the dataset) with descriptive keywords. In designing Wordsmith, we incorporated several elements to engage the player in achieving the goal. We added a colour coded progress timer on each image task and progress bars to track level completion. We also incorporated feed forward alerts to nudge players on when a goal (such as the next level or a new badge) is within reach.

Rules and constraints
Due to the simplicity of Wordsmith as an image labeling game, the rules of Wordsmith merely consist of constraints designed to prevent cheating and input from spam-bots. These constraints consist of a service based English word verification, a restricted-word list to ban common words as players advanced through levels, duplicate checking, and a simple CAPTCHA-like question (e.g. 'What is the current day of the week?') after every 10 image labeling rounds to filter automated processes.

Feedback mechanisms
Feedback consists of information provided to players on their progress and current standing in the game. Providing feedback has been shown to improve player retention and engagement by enhancing intrinsic feelings of accomplishment as players advance. Both social and non-social feedback elements were added to Wordsmith as follows:

1. Leaderboard - showed the hourly scores and level of the top 5 players.
2. Levels - the game consisted of a total of 9 levels from Newbie to Wordsmith. A player's level advancement was a function of how many images were tagged.
3. Badges - were awarded based on the number of images tagged.
4. Feedback Alerts - informed a player how a bonus point or badge was attained and how it can be re-attained.
5. Bonus Points - were awarded when players submitted keywords that matched 1, 2, or 3 known images tags.
6. Treasure Points - were awarded when players got multiples of 10 bonus points.
7. Activities Widget - displayed in realtime, what other players were doing in the game.

Voluntary participation
The final trait was to present the task within the game as what the player chose to do rather than what they were mandated to do. In this regard, Wordsmith supported player freedom in three ways; first, participation of course was purely optional and anonymous, and they could join without registering (filling out a form or providing any personal details). Second, players could terminate at any time; finally, an image skip feature was provided so they could freely skip images they were not interested in.

3.4 Furtherance incentives

In Experiment 2, we introduce "*furtherance incentives*" to Conditions 3 and 4, which is a reward or concession presented to a player when they attempt to exit the game to induce them to stay and play more levels. We selected our incentives based on the SAPS framework presented by [31].

In our experiment, we expanded *Status* from SAPS to encompass the 3 game status elements mentioned by [31], i.e., badges, leaderboard and levels. We interpreted the SAPS incentives as a popup messages presented to the player at the point of attempted exit. Each incentive began with the message: Would you like to tag the next *"target number"* images? If the player had tagged less than 21 images, the "*target number*" of additional images was 5, otherwise it was 11.

The specific messages appended to each incentive is as follows:

- **Badges** - You would automatically be rewarded with The 'Ultimate' Badge. Get upgraded to a shiny new avatar
- **Leaderboard** - You would automatically be advanced on The Global Leaderboard. Get seen globally on the leaderboard
- **Levels** - You would automatically be advanced to The Next Level. Advance to the next level.
- **Access** - You would be given quicker access to Treasure Points. Get more treasure in half the time.
- **Power** - You would be rewarded with the power to View Other Players Tags. Power to see other players image tags.
- **Stuff** - You would be rewarded with a bonus of 5 cents extra. More cash for your effort.

At each point of attempted exit, a player was shown one of the 6 furtherance incentives

$$V = \{badges, leaderboard, levels, access, power, money\}$$

The choice of the incentive to be shown was decided by drawing a random variable $V \backsim U([0,1])$

At the moment of incentive offer, we record the incentive offered, the requested target number, the number of images tagged so far (as start_tags and end_tags) and the current timestamp. We then recorded the player's game state after the incentive was presented i.e., the player could ignore the incentive and exit the game (*state=out*), or the player could go on playing the game (*state=in*).

If the player remains in *state = in*, we keep track of game play (updating the number of images tagged as end_tags) until the player has tagged an additional "target images". At this point, the offered incentive is activated. Players can then transition into state in or out. If a player attempts to exit the game at this point, we do not show any furtherance incentive. However, if the player remains in the game, we continue to keep track of the number of images tagged, and therefore update the value for end_tags.

3.5 Adjusting incentives probabilistically

In the final condition (Experiment 4), we posit that certain furtherance incentives are more effective at different stages of gameplay. To test this hypothesis, we computed a probabilistic model that estimates, at every potential game exit point, the incentive that would maximize the probability of the player remaining in the game. To do this, our probabilsitic model computes *a priori* state transitions at previous attempted exit points, to predict what incentive a player would accept given the number of images they have tagged. This model is computed from:

1. ***the prior probability*** of the incentive given the incentive distribution obtained from the results of the random incentives condition (Experiment 4 Condition 3)
2. ***the likelihood probability*** of the player remaining in the game after tagging the current number of images given a certain offered incentive and
3. ***the likelihood probability*** of the player remaining in the game after tagging a set of images (defined over a numeric range), given a certain offered incentive.

The details of the reasoning approach is detailed in the next few sections.

Our probabilistic reasoning approach to computing the maximum a posteriori incentive given our selected feature (the number of tagged images) is similar to the method of determining the correctness of worker results by [6].

We compute the posterior as the maximum conditional probability of the incentive *v* at a given point *x* using Bayesian inferencing as shown in the equation below:

$$\Pr(v|x, s = in) = \frac{\Pr(x|v, s = in)\Pr(v|s = in)}{\Pr(x|s = in)} \quad (1)$$

where:
v is a potential incentive to be shown to the player.
x is the number of images a player has tagged so far.
s is the state of a player being in or out of the game.
Pr(v|x,s=in) is the posterior of the incentive given the number of images the player has tagged.
Pr(v|s=in) is the prior probability of the incentive i.e., the probability of any player at any given point accepting this incentive.
Pr(x|v,s=in) is the likelihood at the current game point that the player would accept the given incentive.

3.5.1 Definitions

Image point
$x \in X = \{1, ..., N\} where N = 2,200.$
Represents the number of images a player has tagged at the point of an attempted game exit.

Game states
$s \in S = \{out, in\}$
Represents the state a game player is in after attempting to exit the game at an image point.

Incentive
$v \in V = \{badges, leaderboard, levels, access, power, money\}$
Represents the set of incentives from which v is drawn to be presented to the player at the point of attempted exit.

Image Band
$b = (x_i, x_j) = \{b \in \mathbb{R} \mid x_i < b < x_j\}$
Represents a range over image points x_i to x_j over which players exhibit similar exit pattern behaviours.

3.5.2 Prior distributions

Our prior distributions come from the results of random incentives presented to players. We compute the *objective prior* of the incentive *v* as given by the sum and product rule in Bayes Theorem:

$$\Pr(v|s = in) = \frac{\sum_{x=1}^{N} \Pr(s = in|v, x)}{\sum_{x=1}^{N}\sum_{v \in V}^{V} \Pr(s = in|v, x)} \quad (2)$$

This represents the number of players that remained in state *s = in*, at image point *x* after being shown incentive v over all image points x ∈ X, compared with all the players that remained in state *s = in*, at image point *x* after being shown any incentive v in the set of all incentives V over all image points x ∈ X.

As an example, given that 100 players remained in state *s = in* after they were shown any incentive v ∈ V over all image points x ∈ X = {1, ..., 2,200}. If 29 of such players (who remained in the game) were shown incentive v = "*power*", then the prior of the incentive "*power*" over all image points is 29/100 or 29%.

3.5.3 Likelihood distributions

We also compute the likelihood at each image point x, of a worker remaining in state *s = in* given an incentive v. The likelihood represents the conditional probability

$$P(x|v, s = in)$$

at image point x. Our likelihood function was a product of 2 variables: (a) the image point likelihood and (b) the image band likelihood.

Image point likelihood
For each incentive v, we calculate the image point likelihood at point x as:

$$\Pr(x|v, s = in) = \frac{\Pr(s = in, x|v)}{\sum_{s \in S}^{S} \Pr(s, x|v)} \quad (3)$$

This represents how many players remained in state *s = in*, at image point *x* after being shown incentive *v* at image point *x*, compared with all observations of state changes at image point *x* after being shown incentive *v*.

As an example, given that 3 players attempted to stop playing the game after tagging 11 images (image point x = 11) and the 3 players were all shown the incentive v = "*power*". If 2 of the players go on to tag the 12th image, then we calculate the likelihood of a player

remaining in state s=in, at the 11th image, when shown "power" as 2/3.

For image points where we do not have any observed behaviours, i.e., where no player had attempted to exit the game at a certain image point x, we apply the principle of indifference (*principle of maximum entropy*) to accommodate these latent variables.

$$P(x|v, s = in) = 1/N \ \forall \ x \in N = \{1, ..., N\} \tag{4}$$

The variable *x* here represents the image point while N = 2, representing 2 possible states s = {in, out}. Therefore, for unobserved image points,

$$P(x|v, s = in) = P(s = in) = P(s = out) = 0.5 \tag{5}$$

Image band likelihoods

To further accommodate for latent variables and present an expressive picture of how players behave after tagging a certain numbered range of images, we introduced image band likelihoods.

Image Band

$b = (x_i, x_j) = \{b \in \mathbb{R} \mid x_i < b < x_j\}$

Represents a range over image points x_i to x_j over which players exhibit similar exit pattern behaviours.

The image bands b ∈ B were elicited by observations over the results from the randomised incentive condition and they are:

$$B = \{0 - 11, 12 - 60, 61 - 100, 101 - 200, 201 - 2200\}$$

The image band likelihoods were computed on an incentives basis, as such:

$$\sum_{b \in B}^{B} \Pr(s = in, b|v) = 1 \tag{6}$$

For each incentive *v*, we calculate the image band likelihood over band *b* as:

$$\Pr(b|v, s = in) = \frac{\Pr(s = in, b|v)}{\sum_{s \in S}^{S} \Pr(s, b|v)} \tag{7}$$

This represents how many players remained in state *s = in*, within image band *b*, after being shown incentive *v*, compared with all players who remained in state *s = in*, over all image bands b ∈ B, after being shown incentive *v*.

As an example, given that 100 players remained in state *s = in* after they were shown incentive v = "*power*" over all image points x ∈ X = 1, ..., 2,200. If 16 players go on to tag 1 more image within the range of image points x_i, x_j = (12,60) = 12 < b < 60, then the image band likelihood of "12-60" given incentive v = "*power*" is 16/100 or 16%.

3.5.4 Updating the likelihoods

As the experiment runs, we continuously take into account the behaviour of players at each image point. With each new observation at an image point, we recalculate the likelihood of remaining in state *s = in*, at image point *x* after being shown incentive *v*. Therefore, our probabilistic model iteratively updates the likelihoods by constantly learning and taking into account new data based on player interaction. This is of particular importance in filling in revised parameters for the earlier unobserved image points.

As an example, given an image point *x = 20*, where there had been no earlier observations in *experiment 4* of a player exit after being shown incentive v = "*power*". The image point likelihood would be assigned the default of 0.5 (*principle of maximum entropy*), computed as 2 observations with 1 observation at state *s = in*. If a new observation occurs (for any given player) at the image point *x = 20* for incentive v = "*power*" and the player transitions to state *s = out*, the image point likelihood is updated to 3 observations with 1 observation at state *s = in* = 0.33.

3.5.5 Computing the posteriors

Given the incentive prior P(v|s=in) when a player remains in the game, the image point likelihood $\Pr(x|v, s = in)$ and the image band likelihood $\Pr(b|v, s = in)$, we are able to compute the best incentive to offer a player at image point x as the incentive that maximizes the posterior given as

$$\arg\max_{v} \Pr(v|x) = \Pr(j|v, s = in) \Pr(v|s = in) \tag{8}$$

where the joint likelihood of the image point and the image band is given as:

$$\Pr(j|v, s = in) = \Pr(x|v, s = in) \Pr(b|v, s = in)$$

The incentive is then offered to the player and the ensuing state transition is recorded as a new observation point to update the image point likelihood given that incentive.

3.5.6 Algorithms

We now present the algorithm for the image point likelihoods for an incentive and the algorithm for calculating the incentive posteriors at any given image point.

Result: Likelihood P(x|v) = inx/obv
Parameter: v = incentive;
Initialize Latent Variables;
Image Points: x = {1,...,N};
Observations at x: obv = 2;
State = in at x: inx = 1;
for *x in Image Points* **do**
  if *state = in at x* **then**
    obv += 1;
    inx += 1;
  else
    obv += 1;
  end
end

Algorithm 1: Image point likelihoods for incentive v

4. EXPERIMENT DESIGN

This section details the experiments we carried out. We ran 2 experiments. Experiment 1 had 2 conditions - (a) CrowdFlower (non gamified) condition; (b) Wordsmith (gamified) condition. Experiment 2 had 4 conditions - (a) Non gamified condition; (b) Gamified condition (without furtherance incentives); (c) Gamified condition (with random furtherance incentives); (d) Gamified condition (with targeted furtherance incentives).

4.1 Research hypotheses

Our work was centered around 3 potential ways in which gamification can be used to improve paid microtasks:

1. Gamifying paid micro-tasks leads to increased worker engagement, culminating in more work done for less cost.
2. Gamifying paid micro-tasks leads to higher inter-annotator agreement, yielding higher quality results than without.
3. Targeting incentives when a player attempts to quit leads to increased engagement.

Result: Posterior Incentive: $\arg\max_v \Pr(v|x)$
Parameter: x = image point;
Initialize Latent Variables;
Incentive v at x: vx = {};
Posteriors Tracker: pt = {levels = 0, ..., power = 0};
Min Tags: min = 11;
Max Tags: max = 2,200;
for *image tag x from min to max* **do**
  $\Pr(v|x)\ \forall\ v \in V$;
  $\Pr(v|x) = \Pr(v).\Pr(b|v).\Pr(x|v)$;
  Incentive Identifier iid = 0;
  Selected Incentive vx = $\Pr(v|x)$ at iid;
  Update Posteriors Tracker $ptatvx+ = 1$;
  Max Incentive Assignment $mia = \Pr(v) * (max - min)$
  if *pt at vx < mia* **then**
    return vx;
  else
    return vx = $\Pr(v|x)$ at iid + 1;
  end
end

Algorithm 2: Incentive posteriors at image point x

To test these hypothesis, we carried out 2 experiments in image labelling. Workers were shown an image, and asked to assign keywords that describe the image. To test the first two hypotheses, we chose a between-subjects design where the control condition consisted of a standard, non-gamified interface, using CrowdFlower's image labelling job; while the experimental condition, consisted of a gamified interface incorporating several game elements. Both conditions relied on CrowdFlower for worker recruitment, but while workers performed tasks directly within CrowdFlower for the control condition, workers assigned the experimental were redirected to an external game site. Participants in both setups were paid the same amount.

To test the 3rd hypothesis, we carried out 2 additional condition setups on our gamified interface, again with players sourced and redirected from CrowdFlower. In the control condition, workers were shown a randomly-selected incentive to stay when they attempted to leave the game. In the experimental condition, an incentive was shown selected based upon a predictive model constructed from the previous worker's task history. The details of this predictive model was presented earlier in Section 3.5.

4.2 Dataset

For our experiments, we used the ESP game dataset from [26]. This comprises of 100,000 images and about 1.4 million image tags. For each experiment, we selected the images in the dataset which had the highest number of keyword tags associated with it. This was used as a sort of quasi-ground truth, for checking basic tagging quality and assigning bonus points.

4.3 Evaluation metrics

To evaluate worker performance, we measured both the volume of work completed and work quality. To assess the volume of work completed, we simply measured the average number of keywords provided per image. To assess work quality, we used two measures: overlap with the gold standard keywords in the dataset, and a standard measure of inter-annotator agreement from [1] to determine the degree of the pairwise consensus of image labels which were not in the gold standard datset.

For Experiment 2 condition 4, we sought to evaluate the effectiveness of targeted incentives over random ones. To do this, we used as a measure the number of players that tagged at least 1 more image after they were shown each particular incentive. The incentive was shown when they attempted to stop playing the game.

4.4 Experimental conditions

In this section, we summarise both experiments and their conditions in detail.

Experiment 1: *Task: Tag 1 Image with at least 2 keywords; Source dataset size: 200 images; Workers: 600; Payment: $0.02; Platforms: CrowdFlower and Wordsmith*. In the first experiment, workers in either condition were required to tag 1 image with 2 keywords. In Wordsmith, the gamified condition, this corresponded to advancing 1 level into the game. Players in Wordsmith could continue playing the game (tagging more images) after completing the required annotation. There were 200 images in the dataset. Participants were paid 2 cents for the image tagged.

Experiment 2: *Task: Tag 11 images with at least 2 images each; Source dataset size: 2,200; Workers: 600; Payment: $0.10; Platforms: Crowdflower and Wordsmith; Furtherance Incentives:none, random or targeted*. In experiment 2, workers were required to tag 11 images with keywords. However, the dataset size was increased 11 fold (from 200 to 2,200) to allow players to play for longer without seeing repeated images. Intermediate results had shown a number of players tagged the entire dataset of 200 images. This experiment consisted of 4 conditions detailed below. In addition, for conditions 3 and 4, furtherance incentives, defined in Section 3.4 are introduced when players attempt to quit.

Experiment 2 - Condition 1: *Platform: CrowdFlower; Furtherance Incentives: none*. This was a non gamified setup where workers were required to tag 11 images from a dataset of 2,200 images for 10 cents.

Experiment 2 - Condition 2: *Platform: Wordsmith; Furtherance Incentives: none*. In this gamified setup, players were required to tag 11 images from a dataset of 2,200 images for 10 cents. This advanced them 2 levels into the game. The players could continue tagging (playing the game) if they wished.

Experiment 2 - Condition 3: *Platform: Wordsmith; Furtherance Incentives: Random*. Identical to Condition 2, except a random furtherance incentive is presented when a player attempted to exit the game.

Experiment 2 - Condition 4: *Platform: Wordsmith; Furtherance Incentives: Targeted*. Identical to Condition 3, except that the furtherance incentive was selected according to the maximum likelihood of user retention using the probabilistic model presented in Section 3.5.

5. RESULTS

The result of Experiment 1 is summarised in Table 1. The results show that players (participants in the game condition) supplied more keywords than those in the CrowdFlower condition on average (97 per image vs 2), and labeling more images overall (32 per worker vs 1), resulting in an overall yield of 41,206 total keywords vs 1,200 in the control condition. We note that, since the control condition restricted workers to supply up to two keywords for a single image, it is unsurprising that individuals in the control condition provided only two keywords for a single image. However, in both conditions, individuals were rewarded only up through the same amount of work (completing the task of supplying 2 keywords for a single image), and thus the additional work done in the Wordsmith condition was not financially incentivised and done for free. Moreover, compared to the control, the experimental condition yielded significantly more *new keywords*, which we define to be keywords that were not in the original gold standard seed, but

achieved the requisite threshold of inter-annotator agreement. The average inter-annotator agreement, computed as described in 4.3 over all images for the control condition was also much less than that of the experimental condition (5.72% vs 37.7%). The control condition achieved 42.9% coverage of the original gold standard label set, while the experimental condition covered 52.5%.

Table 2 summarises the results for Experiment 2. Again, compared with the CrowdFlower interface, all game conditions saw much greater output, both in terms of labels per image (average 40,510 keywords across game conditions vs 13,200 in the control condition) and number of images tagged (30 images labeled per worker across game conditions vs 11) despite monetary compensation being held constant between conditions (10 cents to complete 11 images with 2 labels each). Examining the game conditions only, conditions 3 and 4 which featured furtherance incentives on exit attempt resulted in players performing more labels on average (31.5 vs 27) than condition 2, which had no furtherance incentives. We note that due to the much larger source dataset of images, the likelihood that two workers would be presented the same image is much lower, resulting overall in noisier inter-rater agreement and lower coverage of gold-standard labels.

To analyse player response to furtherance incentives, Table 4 shows the number of players who responded to each furtherance incentive stimulus type at various levels of play (image bands). To clarify, we considered a player to be *responding to the incentive stimulus* when, upon attempting to quit the game *and* being presented with a furtherance incentive, decided to tag at least 1 extra image prior to exiting. The table shows the number of responses of the number of presentations of each stimuli for each (C3 random and C4 targeted) condition at 5 image image bands, corresponding to the number of images previously tagged when attempting to exit. Comparing randomised to targeted incentive, the results show greater response to furtherance incentives when delivered in the targeted incentive condition (C4) than randomised (C3). In the targeted incentives condition, 77% of players went on to tag at least 1 more image, compared with only 27% in the randomised condition.

With respect to furtherance incentive type, direct comparison is in Table 4 due to the fact that the number of stimulus presentations differ for different types and conditions. We constructed Table 3 to make this comparison further, which simply shows a breakdown, by type, of all successful furtherance incentive stimulus responses. As can be seen, in both C3 (Randomised) and C4 (Targeted), the Power and Money incentives made up the top two successful incentives, with Money comprising the largest share of the targeted successes. We discuss these results in the next section.

Experiment 1		
Metric	CrowdFlower (control)	Wordsmith (experimental)
Total workers	600	423
Total keywords	1,200	41,206
New keywords	111	5,708
Avg. agreement	5.72%	37.7%
Gold keywords	42.92%	52.53%
Mean Imgs/person	1	32
Max Imgs/person	1	200

Table 1: *Experiment 1 Results* - High level results for Experiment 1, comparing number of keywords and images tagged in the gamified (Wordsmith) condition compared to the standard CrowdFlower interface.

Incentive	C3: Randomised	C4: Targeted
Power	26.09%	30.16%
Money	19.65%	46.17%
Leaderboard	16.59%	5.71%
Levels	13.01%	7.34%
Badges	13.04%	5.98%
Access	11.61%	4.35%

Table 3: *Incentive Response Distribution* - Successful furtherance incentives stimuli broken down by type, for both C3 (randomised) and C4 (targeted) conditions.

6. DISCUSSION AND CONCLUSION

In this section, we first briefly re-visit our results in the context of the research hypotheses, discussing limitations in the process. We then discuss implications of our findings to crowdsourcing, and conclude with a summary of ongoing and future work.

The results show support for all three of our research hypotheses. With respect to H1, players in the game condition unilaterally performed more tasks even when they were not explicitly incentivised with monetary reward to do so. In addition, output was of higher quality, showing support for H2, both when measured in terms of diversity (new words with high agreement) and achieved consistent coverage of the gold standard labels than the control condition. In particular, we saw no support for overjustification in these results, which would have been manifest in reduced productivity with the introduction of game elements.

One limitation of our experimental design is that, since the number of contributions in the control interface was clamped while the game condition was not (meaning they could contribute indefinitely), it is not meaningful to quantify the increased volume of work between the control and gamification conditions. However, we can compare quality differences (which showed significant gains), and volume differences among just the game conditions in Experiment 2, when targeted furtherance incentives were shown to yield higher volumes of work than randomised ones (H3).

However, perhaps more significantly, this study demonstrated that even simple furtherance incentivisation methods do work towards getting players to complete more tasks. In all but the Money furtherance incentive condition, such methods worked to increase output at no extra cost. Moreover, we found that among furtherance incentivisation strategies, those that were more social generally fared better than those that were personal; for example, the Power incentive was presented "*You would be rewarded with the power to view other players' tags*", while the Leaderboard incentive promised participants a higher place on the leaderboard, which was visible to everyone. This agrees with previous work in GWAPs such as the the ESP Game, in particular [26]) in which social incentives were shown to be among the most powerful. Most human computation environments, like Mechanical Turk, CrowdFlower and citizen science projects still lack elements that promote social visibility that might improve engagement.

The effectiveness of money as an effective furtherance incentive was somewhat surprising, given the fact that most participants already performed free labour, that is, work beyond the minimum that was asked of them to get their initial reward. Therefore, it could be concluded that these participants were motivated to do this additional work for other reasons. However, when financial reward is re-introduced as a furtherance incentive, it effectively motivated people to complete more work. Further analysis is required to understand to what extent such monetary rewards could compel con-

Experiment 2				
	CrowdFlower (Non-gamified)	Wordsmith (Gamified)		
	C1: No furtherance	C2: No furtherance	C3: Random furtherance	C4: Targeted furtherance
Total workers	600	514	543	454
Total keywords	13,200	35,890	47,418	38,223
New keywords	1,323	4,091	5,435	4,849
Avg. agreement	6.32%	10.90%	10.16%	9.86%
Gold keywords	48.42%	45.02%	41.21%	47.10%
Mean Imgs/person	11	27	33	30
Max. Imgs/person	11	351	501	540

Table 2: *Experiment 2 Results* - High level summary of work output and quality comparing non-gamified (C1) and gamified (C2, C3, C4) conditions.

Image Band	11		12-60		61-100		101-200		201-2,200	
	Random	Targeted	Random	Targeted	Random	Targeted	Random	Targeted	Random	Targeted
Power	26.67% (4/15)	70.97% (22/31)	33.33% (9/27)	74.68% (59/79)	55.55% (5/9)	80.95% (17/21)	25.00% (2/8)	100.00% (5/5)	100.00% (3/3)	61.54% (8/13)
Money	23.53% (4/17)	88.24% (75/85)	34.78% (8/23)	77.57% (83/107)	66.67% (2/3)	100.00% (2/2)	40.00% (2/5)	100.00% (6/6)	0.00% (0/0)	100.00% (5/5)
Leaderboard	21.43% (3/14)	0.00% (0/1)	10.00% (3/30)	70.00% (7/10)	33.33% (1/3)	100.00% (1/1)	70.00% (7/10)	57.14% (12/21)	75.00% (3/4)	100.00% (1/1)
Levels	13.04% (3/23)	40.00% (2/5)	18.18% (6/33)	69.57% (16/23)	75.00% (3/4)	100.00% (3/3)	16.67% (1/6)	100.00% (3/3)	100.00% (2/2)	100.00% (3/3)
Badges	0.00% (0/0)	0.00% (0/1)	22.22% (4/18)	63.63% (7/11)	16.67% (1/6)	100.00% (3/3)	66.67% (4/6)	90% (9/10)	66.67% (2/3)	75.00% (3/4)
Access	11.76% (2/17)	0.00% (0/0)	17.24% (5/29)	33.33% (5/15)	33.33% (3/9)	50.00% (2/4)	20.00% (1/5)	100.00% (3/3)	100.00% (1/1)	100.00% (6/6)

Table 4: *Furtherance Incentive responses (Results of Experiment 2 Condition 3 & 4)* Percentage of players in who responded to each Furtherance Incentive broken down by type (Power, Money, Leaderboard, Levels, Badges and Access), and condition (randomised vs targeted). The number of incentive prompts delivered for each type are listed by the responded percentage.

tinued participation, and the optimal amounts of reward for doing so.

As our experiments only tested one type of crowdsourced task and GWAP, namely image labeling, the results may not necessarily apply to all task types. In particular, tasks that require high cognitive load, require significant time investment or creative thought may not benefit from game mechanics due to their intensive nature. Moreover, those kinds of crowdsourced tasks driven by strong intrinsic motivations (such as citizen science, disaster relief, and so on) are unlikely to substantially benefit from these results because such motivations will probably overshadow the simpler incentives tested here. Moreover, such intrinsic motivation settings have been shown to be more prone to overjustification effects, and thus may result in actually reduced participation. We wish to test whether such effects will become present in such settings in future experiments.

Among our ongoing efforts, we wish to better understand how and why the incentives work in the ways and to the extent that they do. In particular, we believe that furtherance incentives could be more effective if carefully distributed within the game mechanics so that they appear at appropriate intervals when motivation begins to wane, not only after the participant has initiated an attempt to leave.

Second, we wish to the improve the probabilistic model to take into account other aspects of players' performance, task history and demographic, and to understand the ways in estimating appropriate rewards. In particular, we wish to run further experiments to determine whether incentives are more effective for particular demographics than for others, or for workers at particular skill levels or task completion histories.

In addition, we would like to investigate further social effects of furtherance incentives. In this experiment, levels and badges were merely to mark a player's own progress; however, these might be made more effective if such rewards were made visible to other players and seen as a form of status. Such status has been shown to effectively encourage participation in online communities [17] and may translate well to microtask environments as well. Moreover, all of the incentives we applied in this experiment were positive, individual incentives; we next wish to explore the effectiveness of other types of incentives such as positive social incentives (e.g., members of an entire group get a reward), as well as negative incentives both as in-game elments and furtherance incentives. Finally, we would like to understand the span of furtherance incentives, and potential avenues for extending the effects of such incentives in various ways.

In summary, our results have shown that very adding gamification to crowdsourced micro-tasks that already have external in-

centives can improve the quality and quantity of work completed. Our results complement previous work comparing purely gamified and paid crowdsourcing, and extend previous results with a look at multiple kinds of furtherance incentives, combined with reward adaptation. Although we have shown social incentives and supplemental monetary rewards outperform other such incentives, and demonstrated a simple probabilistic model able to outperform a randomised strategy, we believe that we have only begun to understand the relationships that such incentives have on subjective worker experience and sustained engagement in the long term, and plan to continue to pursue such investigations in the future.

7. REFERENCES

[1] P. K. Bhowmick, P. Mitra, and A. Basu. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics, 2008.

[2] R. Burkett. An alternative framework for agent recruitment: From mice to rascls. In *Studies in Intelligence*, volume 57 of *1*, pages 7–17, mar 2013.

[3] R. Dawson and S. Bynghall. *Getting Results from Crowds*. Advanced Human Technologies, 2012.

[4] E. Deci and R. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer, 1985.

[5] E. L. Deci, R. Koestner, and R. M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627, 1999.

[6] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal*, 22(5):665–687, Oct. 2013.

[7] S. Deterding, R. Khaled, L. Nacke, and D. Dixon. Gamification: Toward a definition. In *CHI 2011 Gamification Workshop Proceedings*, pages 12–15, 2011.

[8] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 871–880, New York, NY, USA, 2012. ACM.

[9] C. Harris and P. Srinivasan. Comparing crowd-based, game-based, and machine-based approaches in initial query and query refinement tasks. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. RÃijger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 495–506. Springer Berlin Heidelberg, 2013.

[10] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[11] L. C. Irani and M. Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620. ACM, 2013.

[12] D. Jurgens and R. Navigli. It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics (TACL)*, 2:449–464, 2014.

[13] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk. In *AMCIS'11*, pages –1–1, 2011.

[14] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing: a study on mechanical turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, pages 1–11, 2011.

[15] M. Kearns. Experiments in social computation. *Commun. ACM*, 55(10):56–67, 2012.

[16] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, 2008.

[17] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[18] M. R. Lepper, D. Greene, and R. E. Nisbett. Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and social Psychology*, 28(1):129, 1973.

[19] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In B. Hartman and E. Horvitz, editors, *HCOMP*. AAAI, 2013.

[20] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85, New York, NY, USA, 2009. ACM.

[21] J. McGonigal. *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. Penguin Group , The, 2011.

[22] C. Mellström and M. Johannesson. Crowding out in blood donation: was titmuss right? *Journal of the European Economic Association*, 6(4):845–863, 2008.

[23] P. Michelucci. *Handbook of human computation*. Springer, 2013.

[24] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008.

[25] S. Thaler, E. Simperl, and S. Wölger. An experiment in comparing human-computation techniques. *IEEE Internet Computing*, pages 52–58, 2012.

[26] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[27] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, Aug. 2008.

[28] H. Xie and J. C. Lui. Modeling crowdsourcing systems: Design and analysis of incentive mechanism and rating system. *SIGMETRICS Perform. Eval. Rev.*, 42(2):52–54, Sept. 2014.

[29] D. Yang, G. Xue, X. Fang, and J. Tang. Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, pages 173–184, New York, NY, USA, 2012. ACM.

[30] M. Yin, Y. Chen, and Y. Sun. The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA.*, 2013.

[31] G. Zichermann and C. Cunningham. *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. 2011.