Interacting with Computers Advance Access published October 22, 2015

doi:10.1093/iwc/iwv035

Worker Performance in a Situated Crowdsourcing Market

JORGE GONCALVES[1]*, SIMO HOSIO[1], YONG LIU[2] AND VASSILIS KOSTAKOS[1]

[1]*Department of Computer Science and Engineering, University of Oulu, Pentti kaiteran katu 1, PO Box 4500, FI-90014 Oulu, Finland*
[2]*School of Business, Aalto University, Runebergsgatan 14-16, FI-00100 Helsinki, Finland*
**Corresponding author: jgoncalv@ee.oulu.fi*

We present an empirical study that investigates crowdsourcing performance in a situated market. Unlike online markets, situated crowdsourcing markets consist of workers who become serendipitously available for work in a particular location and context. So far, the literature has lacked a systematic study of task performance and uptake in such markets under varying incentives. In a 3-week field study, we demonstrate that in a situated crowdsourcing market, task uptake and accuracy are generally comparable with online markets. We also show that increasing task rewards in situated crowdsourcing leads to increased task uptake but not accuracy, while decreasing task rewards leads to decreases in both task uptake and accuracy.

RESEARCH HIGHLIGHTS

- We present a 3-week empirical study on worker performance in a situated crowdsourcing market.
- We manipulate task rewards to investigate its effects on performance.
- Increasing task rewards led to increased task uptake but not accuracy.
- Decreasing task rewards led to decreased task uptake and accuracy.
- We compare the performance of our reported tasks and literature using several types of crowdsourcing.

Keywords: user studies; touch screens; empirical studies in HCI; computer supported collaborative work; tablet computers

Editorial Board Member: Michael Muller

Received 10 March 2015; revised 5 April 2015; accepted 13 September 2015

1. INTRODUCTION

This paper is the first study to provide empirical evidence on task performance in a situated crowdsourcing market. The characteristics that set apart situated crowdsourcing from traditional crowdsourcing are location and context, enticing people to physically go to certain locations to complete tasks (as opposed to visiting a website). Situated crowdsourcing is performed using input mechanisms embedded into a physical space (e.g. public displays, tablets). It primarily leverages users' serendipitous availability (Müller *et al.*, 2010), idle time or Shirky's 'cognitive surplus' (Shirky, 2010), in a designated location.

As such, situated crowdsourcing enables geo-fenced and contextually controlled experiments targeting certain populations or communities (Heimerl *et al.*, 2012), leveraging users' local knowledge (Goncalves *et al.*, 2014a) and reaching an untapped source of potential workers (Hosio *et al.*, 2014b). This is in sharp contrast to online crowdsourcing markets that do not always attract workers of a desired background or with a set of skills because the short work duration and small rewards bias the worker demographics (Ross *et al.*, 2010). For example, it is challenging to recruit workers who speak a particular language, live in a given city (Ipeirotis, 2008) or have domain-specific knowledge (Heimerl *et al.*, 2012). This can be detrimental for tasks like the creation of newspaper articles (Alt *et al.*, 2010) or document translation (Zaidan and Callison-Burch, 2011) that require workers within a relevant context.

However, it is yet unclear whether performance in situated crowdsourcing substantially differs from other types of crowdsourcing. A risk with situated technologies is that they are typically in the hands of users, away from the controls of a lab setting and may produce 'noisy' results due to

unpredictable behaviour from users (Hosio *et al.*, 2014a; Schroeter *et al.*, 2012). Furthermore, while they can offer timely contextual information, it is challenging to maintain people's interest and engagement over time (Goncalves *et al.*, 2014b). For these reasons, performance and work quality in a situated crowdsourcing market can be questionable.

In this paper, we systematically measure task accuracy and uptake in a situated crowdsourcing market. Our study was conducted on Bazaar, a situated crowdsourcing market (Hosio *et al.*, 2014b). Previous work has demonstrated that Bazaar has strong market characteristics and follows economic principles, and workers exhibit rationality by changing their behaviour according to price-setting. However, no study has yet investigated task performance in this situated market under varying incentives. We demonstrate that task uptake and accuracy in a situated crowdsourcing market are generally high. Furthermore, our findings suggest that increasing task rewards in situated crowdsourcing will lead to an increase of task uptake but not necessarily an increase in accuracy while decreasing rewards will decrease both task uptake and accuracy.

2. RELATED WORK

2.1. Incentives and performance in crowdsourcing markets

It is important to provide an overview of why people take part as workers in crowdsourcing markets, and what does the theory suggest about their performance in completing tasks. A traditional 'rational' economic approach to eliciting higher quality work is to increase extrinsic motivation, i.e. an employer can increase how much they pay for the completion of a task (Gibbons, 1997). Some evidence from traditional labour markets supports this view: Lazear (2000) found workers to be more productive when they switched from being paid by time to being paid by piece; Hubbard and Palia (1995) found correlations between executive pay and firm performance when markets were allowed to self-regulate.

An experiment by Deci (1975) found a 'crowding out' effect of external motivation: students paid to play with a puzzle later played with it less and reported less interest than those who were not paid to do so. In the workplace, performance-based rewards can be 'alienating' and 'dehumanizing' (Etzioni, 1971). If the reward is not substantial, then performance is likely to be worse than when no reward is offered at all; insufficient monetary rewards can act as a small extrinsic motivation that tends to override the possibly larger effect of the task's likely intrinsic motivation (Gneezy and Rustichini, 2000). Given that crowdsourcing markets such as Mechanical Turk tend to pay very little money and involve relatively low wages (Paolacci *et al.*, 2010), external motivations such as increased pay may have less effect than requesters may desire. Indeed, the research examining the link between financial incentives and performance in Mechanical Turk has generally found a lack of increased quality in worker output (Mason and Watts, 2009). The relationship between price and quality has also had conflicting results in other crowdsourcing applications such as answer markets (Harper *et al.*, 2008). Although paying more can get work done faster, it is unclear if it was performed better.

Another approach to improve work performance could be increasing the intrinsic motivation of the task. Under this view, if workers find the task more engaging, interesting or worth doing in its own right, they may produce higher quality results. Unfortunately, evidence so far has not fully supported this hypothesis. For example, while crowdsourcing tasks framed in a meaningful context motivate individuals to do more, they are no more accurate (Chandler and Kapelner, 2013). On the other hand, the work by Rogstadius *et al.* (2011) suggests that intrinsic motivation has a significant effect on workers' performance.

These contradictory results and a number of other issues that suggest the question of motivating crowd workers has not yet been definitively settled. First, prior studies have methodological problems with self-selection, since workers may see equivalent tasks with different base payment or bonuses being posted either in parallel or serially. Secondly, very few studies besides have looked at the interaction between intrinsic and extrinsic motivations; Mason and Watts (2009) vary financial reward (extrinsic), while Chandler and Kapelner (2013) vary meaningfulness of context (intrinsic) in a fixed diminishing financial reward structure. Finally, the task used in Chandler and Kapelner (2013) resulted in very high performance levels, suggesting a possible ceiling effect on the influence of intrinsic motivation.

In our experiment, we financially reward workers, thus we use extrinsic motivation in a market-driven model rather than intrinsic motivation. This decision was made to increase the external validity of our study, since we wanted to investigate performance under realistic market conditions.

2.2. Crowdsourcing with ubiquitous technologies

Crowdsourcing with ubiquitous technologies is increasingly gaining researchers' attention (Liu *et al.*, 2012; Vukovic and Kumara, 2011), especially on mobile phones. This has allowed researchers to assign tasks to workers, anywhere and anytime. Targeting low-end mobile phones, txtEagle (Eagle, 2009) is a platform for crowdsourcing tasks specific to inhabitants of developing countries. Similar platforms are MobileWorks (Narula *et al.*, 2011) and mClerk (Gupta *et al.*, 2012) that specifically focus on asking users to convert handwritten words to typed text from a variety of vestigial dialects. In a larger project, a mobile crowdsourcing platform called MoneyBee (Govindaraj *et al.*, 2011) was made accessible to mobile phone users in emerging markets through their mobile operators and therefore reaching a higher number of potential workers.

Targeting smartphones, Alt *et al.* (2010) explore location-based crowdsourcing for distributing tasks to workers. They focus on how workers may actively perform real-world tasks for others, such as giving a real-time recommendation for a restaurant, or providing an instant weather report wherever they are. Similarly, Väätäjä *et al.* (2011) report a location-aware crowdsource platform for authoring news articles by requesting photographs or videos of certain events from its workers. Mashhadi and Capra (2011) suggest using contextual information, such as mobility, as a mechanism to ensure the quality of crowdsourced work. Finally, mCrowd (Yan *et al.*, 2009) enables mobile users to utilize sensors on their smartphone to participate and accomplish crowdsourcing tasks, including geolocation-aware image collection, image tagging, road traffic monitoring and others.

An active community has grown around the topic of crowdsourcing measurements and sensing (Liu *et al.*, 2012). This participatory sensing movement is part of the larger concept of 'Citizen Science' (Paulos *et al.*, 2008) that relies on mobilizing large parts of the population to contribute to scientific challenges via crowdsourcing. Often this involves the use of smartphones for collecting data (Burke *et al.*, 2006) or even donating computational resources while one's phone is idle (Arslan *et al.*, 2012).

Despite the appeal of mobile phones, using them for crowdsourcing requires workers' implicit deployment, configuration and use of the device. For example, in SMS-based crowdsourcing, participants need to explicitly sign up for the service, at the cost of a text message exchange. This makes worker recruitment challenging, as a number of steps are necessary before a worker can actually contribute using their device. An alternative approach is to embed input mechanisms (e.g. public displays, tablets) into a physical space and leverage users' serendipitous availability (Müller *et al.*, 2010). This means that, contrary to mobile crowdsourcing, situated crowdsourcing through embedded interfaces does not require any deployment effort from workers (Goncalves *et al.*, 2013; Goncalves *et al.*, 2014c).

In such a deployment, Heimerl *et al.* (2012) reported Umati, which used a vending machine with a touch display for locally relevant tasks, albeit with certain limitations. For example, it was available at a single location only, and it lacked diverse tasks to keep users engaged for long. Goncalves *et al.* (2013) public display crowdsourcing deployment also suffered from the lack of diverse tasks. These findings suggest that task diversity is key to sustaining a situated crowdsourcing market.

3. MARKET DESCRIPTION

Our study was conducted on Bazaar, a situated crowdsourcing market (Hosio *et al.*, 2014b). A full description of the market is beyond the scope of our paper, yet we include all the necessary details relevant to our study and findings. Bazaar has a virtual currency ('HexaCoins') that can be redeemed for goods or cash. It consists of a grid of physical crowdsourcing 'kiosks' coordinated by a single network server that records in detail all user actions and completed tasks. Each kiosk contains an Android tablet with a 10.1 touch-screen, a charger to keep the tablet always on, and uses WiFi to connect to the server. The tablets are set to 'kiosk mode' (Surelock, 2014) to ensure that if the crowdsourcing software is always visible on screen, it recovers from crashes, and unwanted OS functionality (notification bars, etc.) is disabled. The physical buttons of the tablet are obscured by the kiosk's enclosure.

The welcome screen of the kiosks contains a brief introduction to the system, and prompts users to log in or create an account. Registration requires just a username and password. Upon login (Fig. 1), users can work on new tasks and see whether their previous work has been approved. They can also review their HexaCoin balance, transfer them to another user or exchange them for goods/cash.

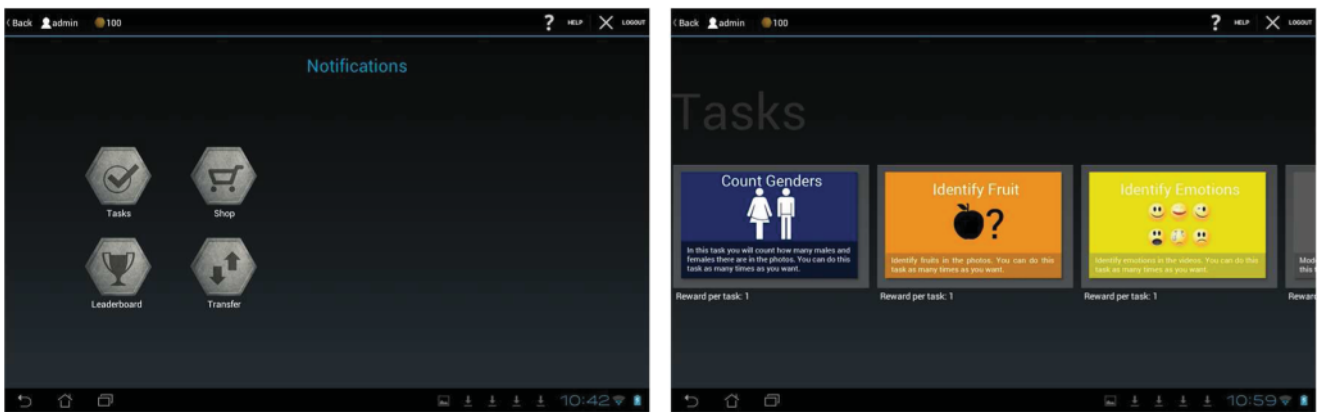

Figure 1. Bazaar's main menu and task menu screens.

3.1. Hexacoins: virtual currency

Bazaar workers are rewarded with HexaCoins which they can in turn exchange for goods or cash. When completing tasks, users receive HexaCoins subject to moderation by administrators or crowd-based moderation. Moderation and rewarding take place in chunks.

The value of HexaCoins is ∼3600 HexaCoins per hour of work. In other words, workers expect to receive one HexaCoin per second of work. This value is influenced by the contextual and cultural factors of the location where the platform is deployed, and therefore these do not follow online prices (e.g. Mechanical Turk). In addition, workers are given 100 free HexaCoins on the first login of each day, to motivate them to return daily and perform more tasks. Users can ultimately exchange HexaCoins for goods, using a rough exchange rate of 360 HexaCoins per 1€. They can obtain cash in 10€ or 25€ packs, and various other goods, including badges, coffee vouchers, movie tickets. Previous work has shown that cash and movie tickets are typically the most popular items on this platform (Hosio *et al.*, 2014b). Workers email the administrators to schedule a pick-up of the items, which is usually preceded by an interview.

4. STUDY

We conducted a 3-week study to investigate workers' performance in Bazaar. During our study, four Bazaar kiosks were active in different buildings of a university campus (Fig. 2). Bazaar is not promoted actively in any way online, only by an A3-sized poster on each of the kiosks. Online promotion is avoided to minimize participation bias.

During this period, we obtained access to the server logs of Bazaar. The server logs all interactions on all kiosks: logins, logouts, starting and ending of performing tasks (time spent), answers for each task. This allowed us to look at task accuracy and task uptake for all tasks and their varying rewards. All users who received goods/cash from Bazaar were interviewed when they picked up their rewards using a standardized interview form.

During our study, there were six different types of tasks available in Bazaar (Table 1) (Hosio *et al.*, 2014b):

(i) *Data categorization*: Categorization and labelling of photographs is a frequently offered crowdsourcing task due to its computational complexity. Workers had to count the number of males and females in a photograph, and another where they had to type the name of the fruit shown in a photograph.
(ii) *Sentiment analysis*: For this task, workers were shown a looping 3-s video of a person's face, and were asked to identify the emotional state of the individual using six response buttons. These buttons correspond to the emotional states that humans can identify quite reliably: anger, happiness, sadness, fear, surprise and disgust (Ekman and Friesen, 1971).
(iii) *Content creation*: For this task, workers had to type a textual description of their surroundings. A worker

Figure 2. Bazaar deployment locations. From left to right: cafeteria (Location 1), next to the main restaurant (Location 2), a lobby with benches (Location 3) and next to a library entrance (Location 4).

Table 1. Summary of number of unique tasks, types, stimuli, worker input and initial reward.

Task category	Unique tasks available	Type	Stimulus	Worker input	Reward (HexaCoins)
Counting genders	373	Data categorization (counting)	Static (images)	Text (numbers)	10
Identifying fruits	370	Data categorization (identification)	Static (images)	Text (short)	10
Identifying emotions	1350	Sentiment analysis	Dynamic (videos)	Multi-choice buttons (6)	5
Describing location	4	Content creation	Text	Text (long)	150
Moderation	Same as number of tasks approved	Content moderation	Static, dynamic, text	Multi-choice buttons (2)	5
Survey	1	Survey	Text	Text and radio buttons	500

could complete this task only once per Bazaar kiosk. This is a task that can greatly benefit from workers' local knowledge (Goncalves *et al.*, 2014a).

(iv) *Content moderation*: For this task, workers had to review other workers' tasks and label them as 'good' or 'bad'. This pool of tasks grew in real-time as workers completed tasks across all Bazaar kiosks. Previous work has shown that crowd-moderation can be a practical approach to quality control (Lampe *et al.*, 2014).

(v) *Survey*: The survey was a one-off task that each worker could complete only once, and only after they had completed 30 other tasks. It contained a set of open-ended questions regarding how they found out about Bazaar, their motivations behind using it, any suggestions of improvements, a standardized System Usability Scale (SUS) and a standardized five-item personality scale (Gosling *et al.*, 2003).

5. RESULTS

5.1. Overall use

As reported in (Hosio *et al.*, 2014b), during the study, we observed sustained use with a total of 194 accounts created, 1067 logins, 75 229 tasks completed (62 602 approved) in 310 114 s (86.1 h) of crowdsourcing effort, and 832 548 HexaCoins generated (Fig. 3). In Fig. 4, we can see the number of logins done by each individual user. As expected with most crowdsourcing studies, there were a number of workers that engaged with our platform purely out of curiosity with about half of them log in in more than once. The most popular task category was moderation ($N = 23\,986$), followed by counting genders ($N = 14\,011$), identifying emotions ($N = 13\,624$) and identifying fruits ($N = 10\,765$). On the other hand, the location description task was completed 138 times and the survey 78 times. A total of 25 transfers were registered (to 10 unique users) worth 14 600 HexaCoins in total. Of the 194 accounts created, 97 (50%) were returning users.

5.2. Accuracy

In Fig. 5, we can see a breakdown of the accuracy of each task. Four out of the five tasks had over 85% accuracy with the highest being the describe location task (∼98%). The identifying emotions task achieved 59% accuracy, suggesting it was a difficult task.

A detailed analysis of the Moderation task reveals similar patterns. This demonstrates that crowd-moderation was mostly effective, except for the Identifying Emotions task (Table 2).

To investigate further why the Identifying Emotions task was performed so poorly, we generated a breakdown of

Figure 3. Cumulative progression of accounts created, logins, tasks approved and time spent crowdsourcing (s) throughout the deployment. (Hosio *et al.*, 2014b)

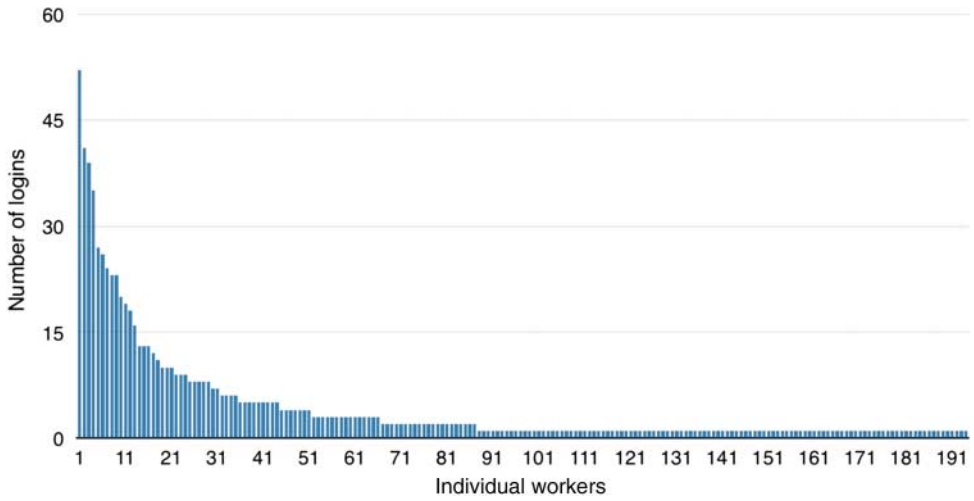

Figure 4. Histogram of frequency of logins for each individual worker.

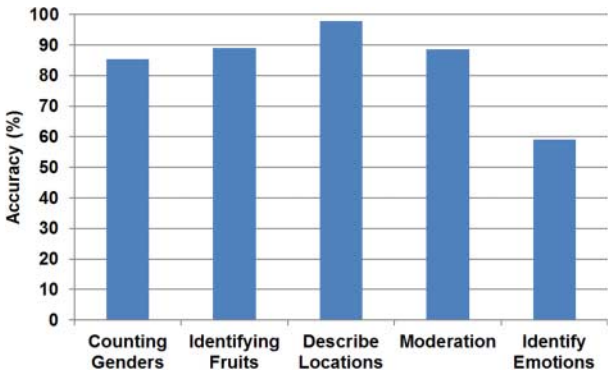

Figure 5. Worker accuracy for each task in Bazaar.

Table 2. Breakdown of correct, incorrect and total number of tasks done in each subcategory of the moderation task as well as its correctness (%).

	Correct	Incorrect	Total	Correctness (%)
Counting genders	3925	717	4642	84.6
Identifying fruits	13 158	1416	14 557	90.4
Describe location	3532	141	3673	96.2
Identifying emotions	636	478	1114	57.1

the answers given by the workers. This breakdown shows that certain emotions were often confused with certain other emotions (Fig. 6). For instance, fear was often confused with surprise, and vice versa. Similarly, anger and disgust were often mixed up in workers' answers.

5.3. Reward manipulation

We experimentally manipulated incentives on a weekly basis to measure the effect on performance. During the first week of deployment, we introduced a reward multiplier, applied to one of the kiosks at a time. For the duration of a whole day, a single kiosk yielded twice (2×) the HexaCoins for each task completed, while all other kiosks operated as usual. We applied this manipulation on four sequential days (Monday–Thursday), each day with the multiplier in a different location. This was done to investigate the performance differences between kiosks that had no multiplier and the one that did considering accuracy and task uptake. In terms of accuracy, the only day the multiplier kiosk performed significantly better than others was Wednesday. Our analysis showed no significant difference between the locations that had a multiplier and those that did not in terms of accuracy: $\chi^2(3) = 0.37$, $P = 0.95$ (Table 3). However, it did have a consistent effect on task uptake: $\chi^2(3) = 1106.21, P < 0.05$ (Table 4).

During Week 2, we modified the rewards of specific task categories, rather than kiosks locations, as follows:

(i) The reward for tasks in the Moderation category was reduced from 5 to 2 HexaCoins. This yielded a 7-fold decrease in task uptake during Week 2 (Fig. 6) and an 8% decrease in accuracy (Fig. 7).
(ii) The reward for tasks in the Identifying Emotions category increased from 5 to 10 HexaCoins. This yielded a 3-fold increase in task uptake during Week 2 (Fig. 6) while the accuracy remained roughly the same (Fig. 7).

During Week 3, we made further manipulations to the rewards per task category as follows:

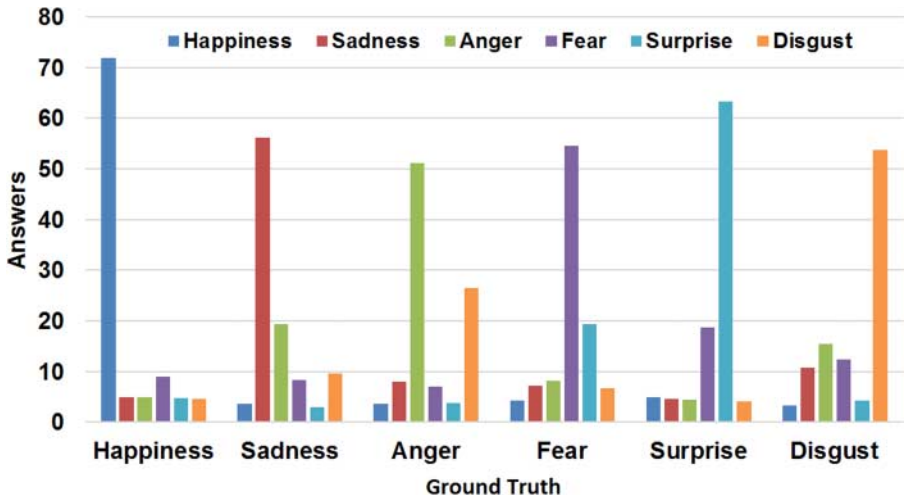

Figure 6. Breakdown of the answers given in the identifying emotions task.

Table 3. Accuracy (%) on each of the locations in the first week of deployment.

	Location 1	Location 2	Location 3	Location 4
Monday	73.3	**78.5**	77.4	84.1
Tuesday	71.9	77.7	83.9	**76.9**
Wednesday	**89.8**	81.0	81.3	82.7
Thursday	91.7	87.4	**87.7**	89.9
Friday	82.1	90.9	88.6	90.1

The underlined bold values correspond to instances where a reward multiplier was present.

Table 4. Task uptake on each of the location in the first week of deployment.

	Location 1	Location 2	Location 3	Location 4
Monday	30	**805**	214	62
Tuesday	281	345	281	**1686**
Wednesday	**6272**	1089	546	393
Thursday	156	2535	**4789**	3850
Friday	178	319	2044	2579

The underlined bold values correspond to instances where a reward multiplier was present.

(i) The reward for tasks in the Identifying Fruits category decreased from 10 to 5 HexaCoins. This yielded a 4-fold decrease in task uptake during Week 3 (Fig. 7) and an 8% decrease in accuracy (Fig. 8).
(ii) The reward for tasks in the Counting Genders category increased from 10 to 15 HexaCoins. This yielded a 10-fold increase in task uptake during Week 3 (Fig. 7) and a 4% decrease in accuracy (Fig. 8).

5.4. Surveys and interviews

As reported in (Hosio *et al.*, 2014b), a total, 78 users completed the survey task (51 male, 27 female). The average age was 23.8 (SD = 4.1). The three most cited reasons for why they used Bazaar were out of curiosity ($N = 41$); to get the rewards illustrated in the posters ($N = 22$); they were recommended by a friend ($N = 14$). When asked where they learned about it we identified two main responses: either the respondents indicated that they just stumbled upon the kiosks at the campus ($N = 55$), or they were informed by their friends about it ($N = 23$).

Analysis of the SUS revealed a score of 81.3 (SD = 10.8) on a scale from 0 to 100. The positive statement with the lowest value for positive was if users would like to use the system frequently ($M = 3.6, \text{SD} = 1.2$). Other values showed that users did not consider the system to be complex ($M = 1.8, \text{SD} = 0.7$), found that it was easy to use ($M = 4.4, \text{SD} = 0.7$), can quickly be learned ($M = 4.4, \text{SD} = 0.7$) and requires no technical support ($M = 1.2, \text{SD} = 0.5$). To provide a fairer grading assignment, we used percentiles like those calculated in Sauro (2011) using a curved grading scale. This means that the SUS score for our system obtained an A grade (above 80.3%).

Finally, 45 workers (26 male, 19 female) of the 194 who created an account in Bazaar purchased prizes and were interviewed during their pick-up of the items. The average age was 23.9 (SD = 3.8). The key findings from the interviews are used to support our discussion.

5.5. Personality influence on performance

A Big-5 personality scale was also part of the survey answered by 78 Bazaar crowd workers. We used this data assess for the potential influence of worker's personality on their performance. The measures we used were as follows:

(i) *Agreeableness*, the tendency to be compassionate, cooperative, trusting and helpful (high score) vs self-interested, suspicious, antagonistic and uncooperative (low score).

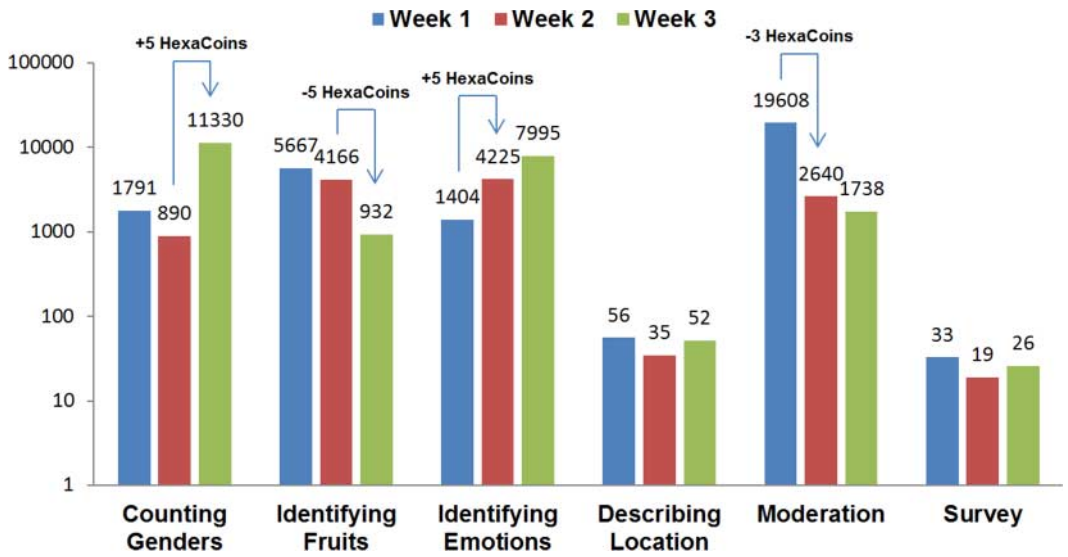

Figure 7. Number of tasks completed per category during each week of deployment (*y*-axis in logarithmic scale). The arrows indicate where a change in reward was done. (Hosio *et al.*, 2014b)

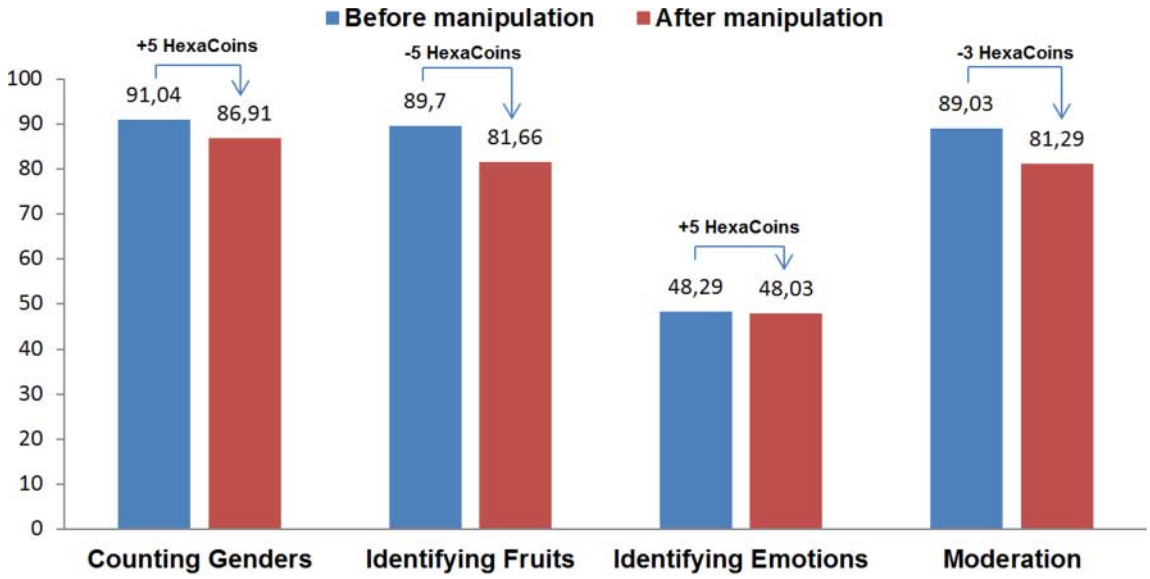

Figure 8. Task accuracy (*y*-axis in %) before and after reward manipulation.

(ii) *Conscientiousness*, the tendency to show self-discipline, act dutifully, be organized, careful and disciplined (high score) vs disorganized, careless and impulsive (low score).
(iii) *Emotional stability*, the tendency to be calm, secure and self-satisfied (high score) vs anxious, insecure and self-pitying (low score).
(iv) *Extraversion*, the tendency to be sociable, fun-loving and affectionate (high score) vs retiring, somber and reserved (low score).
(v) *Openness*, the tendency to be imaginative, independent and interested in variety (high score) vs practical, conforming and interested in routine (low score).

The analysis of the results showed that there was no significant effect of any of the Big-5 personality traits, or any of their combinations, on either task uptake or accuracy.

6. DISCUSSION

Our study is the first in-depth investigation of worker performance in a situated crowdsourcing market. A previous study on Bazaar has focused on its market characteristics such

as price elasticity, adjustment of labour supply using price adjustments, worker preference in terms of financial rewards and moderation of the market (Hosio *et al.*, 2014b). However, they have overlooked worker performance in terms of accuracy and task uptake under varying incentives.

6.1. Worker accuracy and task uptake under varying incentives

The majority of the tasks achieved high levels of accuracy (above 85%) with the only exception being the Identifying Emotion task and the moderation of said task. However, we argue that the subpar worker accuracy for this task was not due to the technology or setting but was a direct cause of its complexity. As seen in Fig. 6, workers had a hard time distinguishing between certain pairs of emotion (e.g. anger/disgust, surprise/fear). One interpretation for these results is that task ambiguity caused workers to arrive to a rushed judgement. Our interviews suggest that this was due to the task's inherent difficulty, and workers decided to move to another task category after completing just a few of the identify emotions tasks as one interviewee stated:

> I liked the emotion task because it was fun with people making funny faces. However, most of the time it was really hard to identify the emotions so I just ended up swapping to other tasks. (P8)

Furthermore, our results show that increasing task rewards in situated crowdsourcing will lead to an increase of task uptake but not necessarily an increase in accuracy (Tables 3 and 4, Figs. 7 and 8). The same phenomenon has also been observed in online crowdsourcing (Chandler and Kapelner, 2013). For instance, during the first week of our deployment, locations with reward multiplier had more tasks completed on average when compared with locations with no reward multiplier (3388 vs 931). At the same time, accuracy between these locations remained fairly equal (83.23 vs 83.37%, respectively).

On the other hand, a decrease in task rewards will lead to both a decrease in task uptake and accuracy (Fig. 8). For example, when we decreased the task reward of the moderation and identifying fruit tasks, we saw a 7- and 4-fold drop in task uptake, respectively, while accuracy also dropped by 8% in both cases.

Therefore, our findings highlight the importance of carefully deciding the rewards given to the tasks so that they remain mostly unchanged. This is of particular importance in situated crowdsourcing as it needs to take into account the cultural and social factors when deciding the rewards. The only scenario where a reward change is advisable, given the same context, is when a task requester needs their task completed quickly and would then increase the reward without sacrificing its accuracy.

Regarding task uptake, we found that Bazaar maintained a task throughput of almost 3000 tasks per day, which is a very high volume compared with previous studies. Previous work has shown that situated crowdsourcing in general tends to have a much greater task uptake when compared with Mechanical Turk. For instance, in a study by Rogstadius *et al.* (2011) non-paid workers took over 45 days to complete 100 tasks while those who were paid (3 or 10 cents) took over 15 and 10 days to complete 200 tasks. Meanwhile, while using the same task Goncalves *et al.* (2013) took 25 days to complete 1200 tasks, without any monetary compensation given. Further, the performance of Umati was compared with that of MTurk, finding that the situated approach with only a single deployed interface was capable of producing 3× more daily labour with over 1000 tasks done daily (Heimerl *et al.*, 2012).

While MTurk studies with high task uptake exist (Lampe *et al.*, 2014), we feel that Bazaar with only four deployed kiosks achieved a workforce throughput that is at least comparable with MTurk, mobile crowdsourcing and previous situated crowdsourcing studies (Table 5).

6.2. Worker abuse and personality

In general, only a handful of workers abused the system by completing tasks in a negligent manner. This type of behaviour can be expected when rewards are given per task rather than per hour (Kittur *et al.*, 2013). However, Bazaar's first-stage moderation and rejection of bad quality work did substantially curb abuse, and in fact we did notice abusive workers eventually produced high-quality work which some workers admitted in the interviews.

> I noticed that after some time my pending tasks were getting deleted instead of awarding me coins. That's when I realised someone was actually checking my answers so I stopped writing nonsense in the identifying fruits task and started answering more seriously. (P22)

Finally, we conducted a test to identify any potential influence of personality traits on worker performance and uptake in situated crowdsourcing. While a prior work (Kazai *et al.*, 2011) suggests that openness significantly relates to accuracy while conscientiousness and agreeableness may also have a positive relation to accuracy, our analysis showed no statistically significant interaction between personality traits and performance. Because only 78 participants completed the survey, and given the field deployment, it is possible that this test did not reliably capture the behavioural traits we were interested in. While our study found no evidence between personality and worker performance, we believe this issue is worthy of a more systematic effort in our future research.

6.3. Crowdsourcing on non-personal devices

A characteristic of situated crowdsourcing that can influence worker performance is that it is performed using non-personal devices as opposed to other means of crowd work. There is a clear distinction between crowdsourcing using one's own personal device (e.g. mobile phone, personal computer)

Table 5. Comparison of performance between the reported tasks (rows 2–5) and literature using several types of crowdsourcing (rows 5–14).

Task description	Stimulus	Worker input	Crowdsourcing type	Input technology	Accuracy (%)	Uptake (tasks per day)	Workers
Counting genders	Image	Text (numbers)	Situated	Public tablet	85	667	194
Identifying fruits	Image	Text (short)	Situated	Public tablet	89	513	194
Identifying emotions	Video	Multi-choice buttons (6)	Situated	Public tablet	59	648	194
Moderation	Image/video	Multi-choice buttons (2)	Situated	Public tablet	89	1142	194
Count cells (Goncalves *et al.*, 2013)	Image	Text (numbers)	Situated	Public display	40–90	48	n/a
Count cells (Rogstadius *et al.*, 2011)	Image	Text (numbers)	Online (MTurk)	Personal computer	66–83	2.2–20	158
Digitize text (Gupta *et al.*, 2012)	Image	Text	Mobile	Personal phone	76–93	1350–2570	221
Digitize text (Narula *et al.*, 2011)	Image	Text	Mobile	Personal phone	89	500	10
Describe current location (Goncalves *et al.*, 2014b)	Current context	Text	Situated	Public display	80–90	200	n/a
Slashdot moderation (Lampe *et al.*, 2014)	Text	Text and buttons	Online (Slashdot)	Personal computer	80	15 665	24 069
Named entity extraction (Finin *et al.*, 2010)	Text	Text	Online (MTurk)	Personal computer	91	800	42
Translation (Eagle, 2009)	Text	Text	Mobile	Personal phone	75	n/a	n/a
Reading task (Kittur *et al.*, 2008)	Text	Likert rating	Online (MTurk)	Personal computer	51	210	58

versus a non-personal device that is embedded in the urban space (e.g. kiosks). A key affordance of performing tasks using your own personal device when compared with non-personal devices is that it can be done in the comfort of one's home, and using familiar technology. However, workers of mobile crowdsourcing need to explicitly sign up for the service (potentially at the cost of a text message exchange) and normally they cannot really control when they receive requests on their phones unless the system allows them to specify when they wish not to be disturbed (Church *et al.*, 2014). This means that task requests may come at inopportune times and lead to disinterest over time (Gupta *et al.*, 2012). Meanwhile, online crowdsourcing gives workers much more flexibility and possibly the best interaction experience out of all forms of crowdsourcing. However, it requires workers to actively look for and sign up to crowdsourcing markets which limits the type of workers that perform tasks.

On the other hand, Müller *et al.* argue that situated technologies do not invite people for a single reason, but users come across and start to use them with no clear motives in mind (Müller *et al.*, 2010). Therefore, they reach users that could otherwise be hard or borderline impossible to reach and ultimately have at that moment free time to spare. As noted by the vast majority of interviewees, most workers of Bazaar were completely new to crowdsourcing, and admitted to have never used any of the popular crowdsourcing markets such as Amazon's Mechanical Turk and CrowdFlower. This strongly indicates the potential of situated crowdsourcing to reach untapped populations of workers and enable high task uptake. Thus, while crowdsourcing with situated technologies is still just an emerging opportunity, research in the area is encouraging and motivates further exploration. Similar findings have been demonstrated in the past, in the context of bridging citizens and city officials through situated technologies (Hosio *et al.*, 2015). In that study, users were able to contribute serendipitously with low effort. This further suggests that situated technologies can appeal to a whole new user base or crowd worker.

Finally, a previous work suggests that crowdsourcing deployments that leverage situated technologies should be designed for 'loners', because groups of people may exhibit non-serious performance when completing crowdsourcing tasks (Goncalves *et al.*, 2013). However, this may be difficult to achieve when using bigger public displays for crowdsourcing as people feel a certain awkwardness and external pressure when interacting alone in public, where passersby can observe them using the display (Brignull and Rogers, 2003). This often leads to displays being used simultaneously by multiple users (most likely friends) (Hosio *et al.*, 2014a) potentially leading to a dip in the quality of crowdsourcing contributions. In our study, through the use of tablets Bazaar makes performing crowdsourcing tasks with situated technologies a more 'personal' experience. This is highlighted in our interviews where workers reported being comfortable performing the

tasks publicly on account of their body occluding the screen and that they would only approach the kiosks when alone.

> My body pretty much prevented anyone else from seeing what I was doing so it was not problem at all for me. (P9)

> Did not even think about feeling self-conscious doing this in public. Nothing different in here, it's just the same as fiddling with my phone. Also, as it is just there already physically, it is the same as the computer terminals at university. (P10)

6.4. Limitations

We acknowledge certain limitations in this study. We encountered run-time problems particularly with WiFi connectivity, leading to suboptimal user experience at times. This is, however, to be expected with any real-world deployment, and the outages usually lasted just a few minutes. The length and magnitude of the deployment, we feel, counterbalances the issue. Finally, cultural issues were not investigated, which could affect the acceptability and performance of situated crowdsourcing.

7. CONCLUSION

This study investigates worker performance in the situated crowdsourcing market Bazaar. Particularly, we look at levels of task uptake and accuracy across different tasks, and fluctuations caused by manipulating incentives. Our results show that task uptake is generally high compared with previous crowdsourcing studies, while accuracy was also high except for one difficult task on sentiment analysis.

Furthermore, through manipulating the rewards for different locations and tasks, we demonstrate that while increasing rewards will yield higher uptake of tasks, it will not necessarily lead to an increase in accuracy. On the other hand, decreasing rewards will ultimately lead to a decrease in uptake and accuracy. These findings have obvious implications for price-setting on a situated crowdsourcing market because task requesters should carefully deliberate the reward to avoid making changes throughout the task's life cycle. In general, increasing the price is easier than decreasing the price.

In addition, we compare the performance between the reported tasks and literature using several types of crowdsourcing. We show that Bazaar with only four deployed kiosks achieved a workforce throughput that is at least comparable with MTurk, mobile crowdsourcing and previous situated crowdsourcing studies. Finally, we discuss the impact on performance given the characteristics of crowdsourcing on non-personal devices and of the situated crowdsourcing workforce. By making the situated crowdsourcing experience more 'personal' by using tablet instead of more public input mechanisms (e.g. public displays), our platform invited more 'loners' which have been shown to be the ideal worker in situated crowdsourcing (Goncalves *et al.*, 2013).

FUNDING

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 285062-iCYCLE, 286386-CPDSS, 285459-iSCIENCE), and the European Commission (Grants PCIG11-GA-2012-322138 and 645706-GRAGE).

REFERENCES

Alt, F., Shirazi, A.S., Schmidt, A., Kramer, U. and Nawaz, Z. (2010) Location-based Crowdsourcing: Extending Crowdsourcing to the Real World. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI'10)*, pp. 13–22. ACM.

Arslan, M.Y., Singh, I., Singh, S., Madhyastha, H.V., Sundaresan, K. and Krishnamurthy, S.V. (2012) Computing while Charging: Building a Distributed Computing Infrastructure using Smartphones. In *Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies (CoNEXT'12)*, pp. 193–204. ACM.

Brignull, H. and Rogers, Y. (2003) Enticing People to Interact with Large Public Displays in Public Spaces. In *Proceedings of INTERACT*, pp. 17–24. IOS Press.

Burke, J.A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S. and Srivastava, M.B. (2006) Participatory Sensing. In First Workshop on World-Sensor-Web: Mobile Device Centric Sensory Networks and Applications at Sensys'06. ACM.

Chandler, D. and Kapelner, A. (2013) Breaking monotony with meaning: motivation in crowdsourcing markets. J. Econ. Behav. Organ., 90, 123–133.

Church, K., Cherubini, M. and Oliver, N. (2014) A large-scale study of daily information needs captured in-situ. ACM Trans. Comput.-Hum. Interact., 21, 1–46.

Deci, E. (1975) *Intrinsic Motivation*. Plenum Press, New York.

Eagle, N. (2009) txtEagle: Mobile Crowdsourcing. In Proceedings of the 3rd International Conference on Internationalization, Design and Global Development: Held as Part of HCI International 2009 (IDGD'09), pp. 447–456. Springer, Berlin, Heidelberg.

Ekman, P. and Friesen, W.V. (1971) Constants across cultures in the face and emotion. J. Pers. Soc. Psychol., 17, 124.

Etzioni, A. (1971) *Modern Organizations*. Prentice-Hall, Englewood Cliffs, NJ.

Finin, T., Murnane, W., Karandikar, A. and Keller, N. (2010) Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of Human Language Technologies Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 80–88. ACL.

Gibbons, R. (1997) Incentives and Careers in Organizations. In Kreps, D. and Wallis (eds), Advances in Economic Theory and Econometrics, Vol. II. Cambridge University Press.

Gneezy, U. and Rustichini, A. (2000) Pay enough or don't pay at all. Q. J. Econ., 115, 791–810.

Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H. and Kostakos, V. (2013) Crowdsourcing on the Spot: Altruistic use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of UbiComp'13*, pp. 753–762. ACM.

Goncalves, J., Hosio, S., Ferreira, D. and Kostakos, V. (2014a) Game of Words: Taggeing Places through Crowdsourcing on Public Displays. In *Proceedings of DIS*'14, pp. 705–714. ACM.

Goncalves, J., Kostakos, V., Karapanos, E., Barreto, M., Camacho, T., Tomasic, A. and Zimmerman, J. (2014b) Citizen motivation on the go: the role of psychological empowerment. Interact. Comput., 26, 196–207.

Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G. and Kostakos, V. (2014c) Projective Testing of Diurnal Collective Emotion. In *Proceedings of UbiComp'14*, pp. 487–497. ACM.

Gosling, S. D., Rentfrow, P. J. and Swann, W. B. (2003) A very brief measure of the Big-Five personality domains. J. Res. Pers., 37, 504–528.

Govindaraj, D., KVM, N., Nandi, A., Narlikar, G. and Poosala, V. (2011) MoneyBee: Towards enabling a ubiquitous, efficient, and easy-to-use mobile crowd sourcing service in the emerging market. Bell Labs Technical Journal, 15, 79–92.

Gupta, A., Thies, W., Cutrell, E. and Balakrishnan, R. (2012) mClerk: Enabling Mobile Crowdsourcing in Developing Regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, pp. 1843–1852. ACM.

Harper, F.M., Raban, D., Rafaeli, S. and Konstan, J.A. (2008) Predictors of Answer Quality in Online Q&A Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*, pp. 865–874. ACM.

Heimerl, K., Gawalt, B., Chen, K., Parikh, T. and Hartmann, B. (2012) CommunitySourcing: Engaging Local Crowds to Perform Expert Work via Physical Kiosks. In *Proceedings of CHI'12*, pp. 1539–1548. ACM.

Hosio, S., Goncalves, J., Kostakos, V. and Riekki, J. (2015) Crowdsourcing Public Opinion using Urban Pervasive Technologies: Lessons from Real-Life Experiments in Oulu. Policy & Internet, 7, 203–222. Wiley Online Library.

Hosio, S., Goncalves, J., Kostakos, V. and Riekki, J. (2014a) Exploring Civic Engagement on Public Displays. In Saeed, S. (ed.), User-Centric Technology Design for Nonprofit and Civic Engagements, pp. 91–111. Springer International Publishing.

Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D. and Kostakos, V. (2014b) Situated Crowdsourcing Using a Market Model. In *Proceedings of UIST'14*, pp. 55–64.

Hubbard, R.G. and Palia, D. (1995) Executive pay and performance Evidence from the US banking industry. J. Financ. Econ., 39, 105–130.

Ipeirotis, P. (2008) Mechanical Turk: The Demographics. In A Computer Scientist in a Business School. http://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf (accessed 5 March 2015).

Kazai, G., Kamps, J. and Milic-Frayling, N. (2011) Worker Types and Personality Traits in Crowdsourcing Relevance Labels. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 1941–1944. ACM.

Kittur, A., Chi, E.H. and Suh, B. (2008) Crowdsourcing user Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*, pp. 453–456. ACM.

Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. and Horton, J. (2013) The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13)*, pp. 1301–1318. ACM.

Lampe, C., Zube, P., Lee, J., Park, C.H. and Johnston, E. (2014) Crowdsourcing civility: a natural experiment examining the effects of distributed moderation in online forums. Gov. Inf. Q., 31, 317–326.

Lazear, E.P. (2000) Performance pay and productivity. Am. Econ. Rev., 90, 1346–1361.

Liu, Y., Lehdonvirta, V., Alexandrova, T. and Nakajima, T. (2012) Drawing on mobile crowds via social media. Multimedia Syst., 18, 53–67.

Mashhadi, A.J. and Capra, L. (2011) Quality Control for Real-time Ubiquitous Crowdsourcing. In *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsouring*, pp. 5–8. ACM.

Mason, W. and Watts, D.J. (2009) Financial Incentives and the 'Performance of Crowds'. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP'09)*, pp. 77–85. ACM.

Müller, J., Alt, F., Michelis, D. and Schmidt, A. (2010) Requirements and Design Space for Interactive Public Displays. In *Proceedings of the International Conference on Multimedia (MM'10)*, pp. 1285–1294. ACM.

Narula, P., Gutheim, P., Rolnitzky, D., Kulkarni, A. and Hartmann, B. (2011) MobileWorks: A Mobile

Crowdsourcing Platform for Workers at the Bottom of the Pyramid. In *AAAI Workshop on Human Computation*, pp. 121–123. AAAI.

Paolacci, G., Chandler, J. and Ipeirotis, P. (2010) Running experiments on amazon mechanical Turk. Judgment Decis. Mak., 5, 411–419.

Paulos, E., Honicky, R.J. and Hooker, B. (2008) Citizen Science: Enabling Participatory Urbanism. In Foth, M. (ed.), Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, pp. 414–436. IGI Global, Hershey, PA.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M. (2011) An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *AAAI Conference on Weblogs and Social Media*, pp. 321–328.

Ross, J., Irani, L., Silberman, M., Zaldivar, A. and Tomlinson, B. (2010) Who are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pp. 2863–2872. ACM.

Sauro, J. (2011) *A Practical Guide to the System Usability Scale (SUS): Background, Benchmarks & Best Practices*. Measuring Usability LLC.

Schroeter, R., Foth, M. and Satchell, C. (2012) People, Content, Location: Sweet Spotting Urban Screens for Situated Engagement. In *Proceedings of the Designing Interactive Systems Conference (DIS'12)*, pp. 146–155. ACM.

Shirky, C. (2010) Cognitive Surplus: How Technology Makes Consumers into Collaborators. Penguin.

SureLock. (2014) http://www.42gears.com/surelock/ (accessed 3 March 2015).

Väätäjä, H., Vainio, T., Sirkkunen, E. and Salo, K. (2011) Crowdsourced News Reporting: Supporting News Content Creation with Mobile Phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'11)*, pp. 435–444. ACM.

Vukovic, M. and Kumara, S. (2011) Second International Workshop on Ubiquitous Crowdsourcing: Towards a Platform for Crowd Computing. In *Proceedings of the 13th International Conference on Ubiquitous Computing Adjunct (UbiComp'11)*, pp. 617–618. ACM.

Yan, T., Marzilli, M., Holmes, R., Ganesan, D. and Corner, M. (2009) mCrowd: A Platform for Mobile Crowdsourcing. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pp. 347–348. ACM.

Zaidan, O.F. and Callison-Burch, C. (2011) Crowdsourcing Translation: Professional Quality from Non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1220–1229. ACL.