# Experiments on Motivational Feedback for Crowdsourced Workers

**Tak Yeon Lee[*], Casey Dugan[+], Werner Geyer[+], Tristan Ratchford[+], Jamie Rasmussen[+], N. Sadat Shami[#] and Stela Lupushor[#]**

[*]Human-Computer Interaction Lab, Computer Science Dept., University of Maryland, College Park, MD 20742 USA, reflect9@gmail.com

[+]IBM T.J. Watson Research, Cambridge, Massachusetts, USA, {cadugan, werner.geyer, tratch, jrasmus} @us.ibm.com

[#]IBM, Armonk, New York, USA, {sadat, stela.lupushor}@us.ibm.com

**Abstract**

This paper examines the relationship between motivational design and its longitudinal effects on crowdsourcing systems. In the context of a company internal web site that crowdsources the identification of Twitter accounts owned by company employees, we designed and investigated the effects of various motivational features including individual / social achievements and gamification. Our 6-month experiment with 437 users allowed us to compare the features in terms of both quantity and quality of the work produced by participants over time. While we found that gamification can increase workers' motivation overall, the combination of motivational features also matters. Specifically, gamified social achievement is the best performing design over a longer period of time. Mixing individual and social achievements turns out to be less effective and can even encourage users to game the system.

## Introduction

Prior economic, sociological and psychological research has identified motivational factors in crowdsourcing applications such as Wikipedia or Mechanical Turk (Kaufmann, Schulze, and Veit 2011). "Crowdsourcing", which outsources a task to an undefined network of laborers using a type of "open call" (Howe 2006), is particularly suited to take advantage of relevant motivational factors to improve the performance of workers. While prior research provides theoretical guidance on motivational factors, designing a successful system remains challenging for several reasons. First of all, finding causal relationships between motivational factors and their effectiveness demands experimental study. For instance, while a traditional economic approach believes financial reward leads to higher quality work (Gibbons 1996), many experimental studies indicate that the combinatorial effect of motivational factors often determines the quantity and quality of workers' output (Rogstadius et al. 2011).

Crowdsourced workers experience motivational factors through task properties and UI designs. In order to employ a specific motivational factor, a designer must be able to choose the right combination of techniques, and carefully design them. For instance, *playfulness*, a commonly used motivational factor can be created with various techniques such as *peer-competition*, *timed-response rule* or a *leaderboard* (von Ahn, and Dabbish 2008). Moreover, a technique does not necessarily have one-to-one relationships with a single motivational factor, making it hard to control for which factor influences workers' motivation. For example, a *leaderboard* implements multiple motivational factors, such as *playfulness* and *social reputation*.

In this paper, we present the results of a controlled experiment that explored the design space of an achievement feedback UI and compared the combinatorial effect of the motivational factors involved. The hypotheses for our experiment were as follows:

**H1.** Workers with no feedback will contribute the least.
**H2.** Providing more motivational elements will increase worker's motivation.
**H3.** Gamification will increase the amount of contribution.
**H4.** Gamification will lower the quality of contributions by encouraging workers to cheat.

The experiment was conducted with users of IBMersWhoTweet, a web application we developed and deployed internally at IBM that aggregates tweets of its employees. The site crowdsources the identification of Twitter accounts (owned by company employees) by assigning identification tasks to workers (other employees). The experiment was run for 6 months in IBM

with over 437 voluntary participants who have identified 2,427 employees, 368 groups, 225 ex-employees, and 1,134 irrelevant Twitter accounts. Our findings indicate that the motivational factors significantly effected workers' behavior. We found gamification to be an effective technique to increase motivation; however, adding motivational elements requires extra care because it can potentially decrease participation or encourage users to game the system. Our research contributes to the understanding of the relationships between various feedback UI designs and the quality/quantity of user participation on IBMersWhoTweet. Also, based on our results, we present guidelines for designers who build crowdsourcing systems that employ feedback as a main motivational feature.

In the next section we discuss motivational factors and elements related to crowdsourcing. Then we introduce IBMersWhoTweet, the web site that our crowdsourcing experiment runs on, and the task given to workers. The following sections describe our experimental method including the 7 motivational settings we compare, and statistical tests and charts for finding causal relationships between feedback UI designs and the quality/quantity of user participation. Finally we conclude with a discussion of our findings and future work in the last section.

## Motivational Structure of Crowdsourcing Systems

A large body of prior work offers theoretical guidance on the motivational factors that make people participate. Deci and Ryan (1985) defined a model of motivational factors classified into either intrinsic or extrinsic motivations, and Kaufmann et al (2011) elaborated on the model for crowdsourced workers. According to the model, intrinsic motivations (e.g. fun, autonomy, reputation) are driven by personal interest and internal emotions in the task itself, while extrinsic motivations (e.g. money, learning, forcedness, implicitness) are influenced by the context of the work. Quinn and Bederson (2011) identified pay, altruism, enjoyment, reputation, and implicit work as motivation constructs in Human Computation systems, which was coined by von Ahn, and Dabbish (2004). Kaufmann et al. (2011) and Ipeirotis (2010) analyze what motivates workers in paid crowdsourcing environments (e.g. Amazon Mechanical Turk; MTurk). Brabham (2008; 2010) surveys the motivation for submitting photos and T-shirt designs online. BJ Fogg's behavior model (Fogg 2009) extends the above motivational theory by adding two more factors: *ability* and *trigger*. According to the model, the worker must be sufficiently motivated, have the *ability* to perform the task, and be *triggered* to perform the behavior.

Controlled experiments focus on how motivational factors influence the quality and quantity of workers' outcomes. Shaw et al. (2011) conducted a study on MTurk that compares the effect of 14 different "social" and "financial" *incentive schemes* in the form of textual instruction. Chandler and Kapelner (2010) observed the effect of *meaningfulness* of a task. Kinnaird et al. (2012) tested whether *workflow transparency* increased worker's volunteerism. It is noteworthy that most controlled experiments on crowdsourcing were done on MTurk where all the tasks are short-term and financial incentive is the strongest motivational factor. On the contrary, IBMersWhoTweet is an internal enterprise service where monetary rewards could not be leveraged to spur worker motivation in completing tasks. This makes it more similar to GWAPs (von Ahn, and Dabbish 2008), MovieLens (Cosley et al. 2003) and StackOverflow[1].

### Motivational Techniques

In this section, we list techniques commonly used for implementing motivational factors.

**Background & Instruction** can frame its task as more meaningful, social, or enjoyable. For example, telling workers that they are finding tumor cells for curing cancer is likely to increase both the quantity and the quality of participation, compared to simply giving them the task (Chandler, and Horton 2011). This technique relies mostly on intrinsic motivations, and requires extra care when being coupled together with extrinsic motivations. For example, too much extrinsic motivation (e.g. financial reward) can undermine intrinsic motivations such as altruism or playfulness (Mason and Watts 2009; Gneezy and Rustichini 2000; Heyman and Ariely 2004).

**Trigger** relates to the when and where of showing a specific task to potential workers. An effective trigger can significantly increase the quantity and quality of participation by taking advantage of moments when the worker's motivation for the task is high, or choosing tasks that fit the worker's current interests (Fogg 2009). Despite the importance of triggers, many online labor markets (i.e. MTurk) provide little flexibility in using custom triggers. Chandler and Horton (2011) showed that tasks placed at focal positions are more preferred than tasks at non-focal positions. For socio-technical platforms such as forums and wikis, triggers are important means to motivate *readers* (passive consumers) to become *leaders* (active contributors) (Preece, and Shneiderman 2009). reChapcha and GWAPs (von Ahn, and Dabbish 2004) embed triggers in the middle of other activities, so that tasks become implicit work for participants.

[1] http://stackoverflow.com/

**Incentive schemes** are an important part of any system relying on extrinsic motivations (i.e. MTurk). Shaw et al. conducted a study on MTurk that compares the effect of 14 different "social", "financial", and "hybrid" incentive schemes in the form of textual instruction, however, not every design choice makes a significant impact on worker's performance (Shaw, Horton, and Chen 2011). Paying more will increase the quantity but not necessarily the quality of the work and can potentially undermine intrinsic motivations (Mason et al. 2009; Gneezy et al 2000; Heyman, and Ariely 2004). Incentive schemes can activate various motivational factors such as *collectivism* (Shaw et al 2011), *playfulness* (von Ahn et al. 2008) by giving incentives related to them.

**Tools & Environments** define how workers do the task. Although they are commonly designed just to lower the required ability and to increase the productivity, workers on a crowdsourcing platform can find intrinsic motivations (e.g. *self-achievement* and *playfulness*) while using the tools and the environments. Luis von Ahn has explored various ways to create extra playfulness with *pairwise competition* and *limited-time response* (von Ahn 2006; von Ahn 2008). Hackman and Oldham (1980) suggested that *task autonomy* makes workers experience responsibility for outcomes of their work, thus gaining high intrinsic motivation over time.

**Feedback** is a common way of giving rewards for tasks done. Gamification techniques (i.e. badges, levels, progress bars, leader boards, virtual currency etc.) have been applied to a wide range of social platforms (e.g. online learning[2], question-and-answer[1], and restaurant reviews[3]) and can create extra playfulness (Deterding, Dixon, Khaled, and Nacke 2011; von Ahn, and Dabbish 2008). Feedback on the quality of work (i.e. self-assessment or external review) makes workers put forth more effort, and thus can yield higher quality results (Dow et al 2012). Kinnard et al. (2012) showed that showing the worker's contribution to the entire process increases volunteerism.

Among the techniques above, we chose a Feedback element and explored its design space for versatility. Common feedback messages, such as the number of tasks done, achievement badges, and rankings, are applicable to a wide range of crowdsourced tasks. Also, implementing various motivational factors (e.g. peer-competition, self-achievement and collectivism) using feedback requires relatively small design changes within the feedback UI element, which reduces the potential bias of our experiment.

## IBMersWhoTweet

IBMersWhoTweet is a novel web application we developed and deployed internally at IBM. The main goal of the application is aggregating IBMers' tweets, and showing those in a customized tweet stream. This goal addresses the needs of three different user groups. First, employees who want to listen to what their peers are saying on external social media find the app an interesting information channel. Second, for employees who want to share their opinions with other employees, IBMersWhoTweet is a useful tool as well. Last, from an

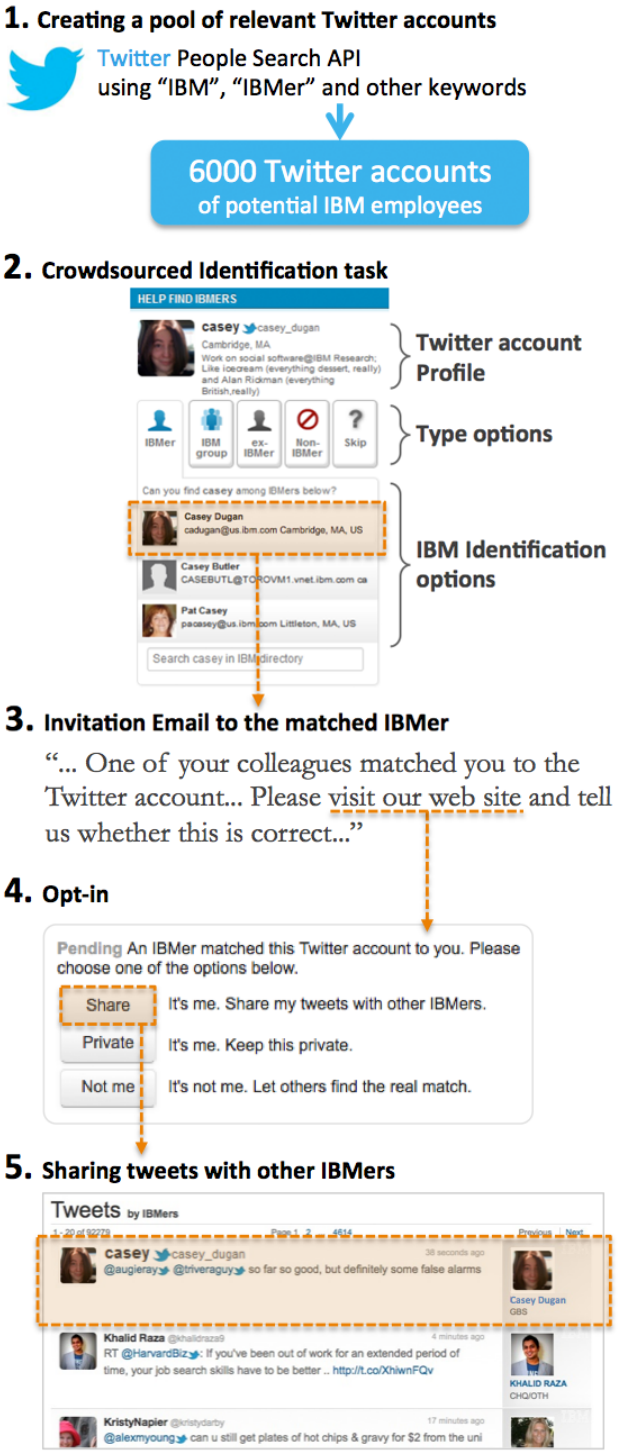


*Figure 1. The Process of Adding IBM Employee Twitter Accounts.*

[2] https//www.khanacademy.org/
[3] http://www.yelp.com

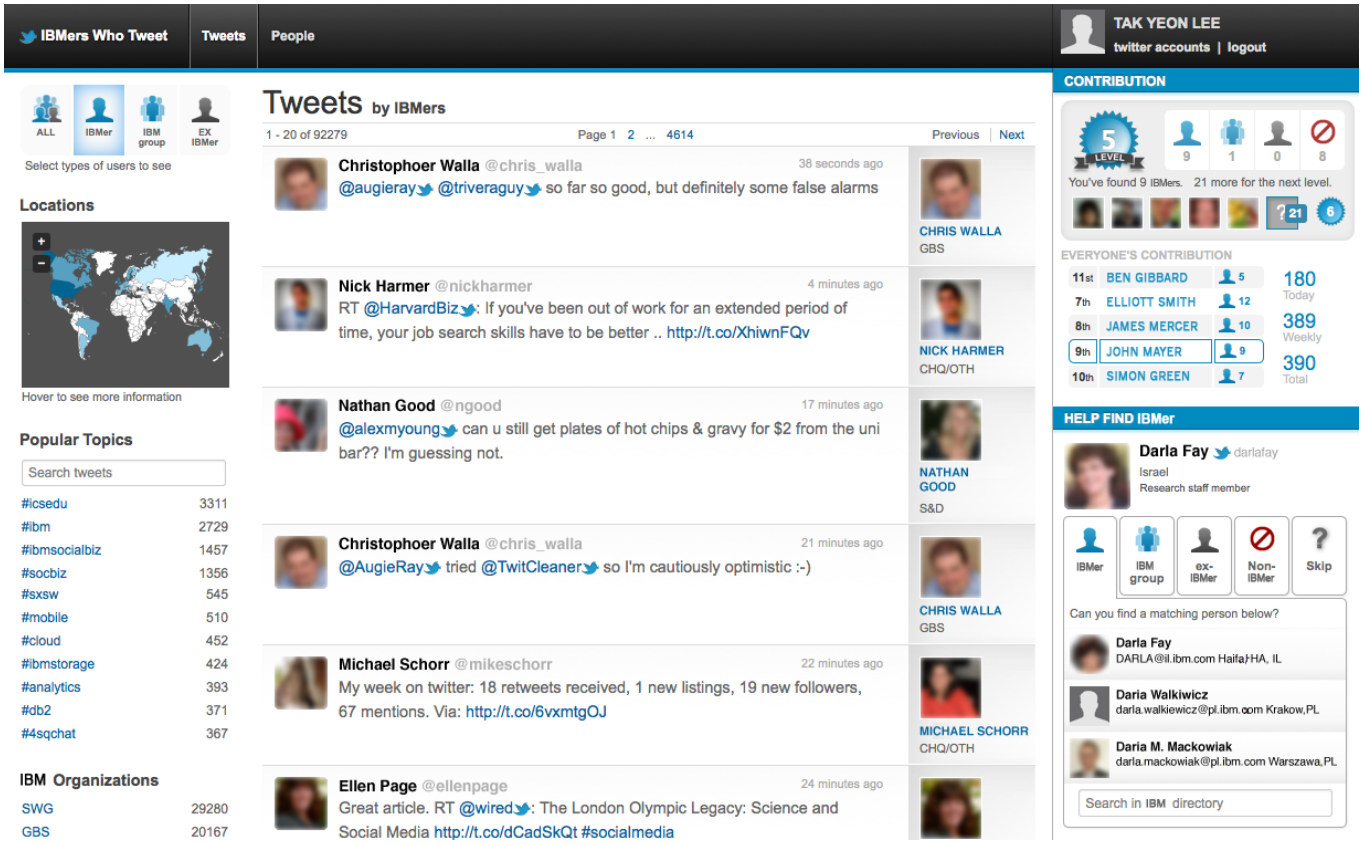


*Figure 2. IBMersWhoTweet UI**; Tweets** of identified IBMers are listed in the middle column. Filtering options for tweets are shown on the left. The panel on the right side has features for crowdsourcing tasks; (Bottom Right) **Help Find IBMers** widget shows a potential employee Twitter account and options for workers to identify the employee who owns it. (Top Right) **Contribution** panel shows worker's achievements. Profile photos and names are anonymized in the screenshot.*

enterprise point of view, marketers, communication professionals and human resource managers can use it to understand the "voice of your company".

The privacy of identified employees, and handling misclassifications are significant issues of this kind of online participation. We briefly touch on them in the Identification Process Section. However, a full discussion of these is beyond the scope of this paper. To our knowledge, these issues did not interfere with our experimental settings.

## Identification Process

### Step 1. Creating a pool of relevant Twitter accounts

The identification process (*Figure 1*) begins by creating a pool of potentially relevant Twitter accounts that may be owned by IBM employees or related to IBM in some way. This pool is seeded with calls made to the Twitter user search API[4], which returns users matching a given keyword, in this case "IBM." Given that this API has a technical limit of returning only the first 1,000 results, we expand the pool of potential accounts in a number of ways. For example, we can pull in users appearing on lists, created by that initial pool of users, which include "IBM" in the title. Additionally, we can consider those *followed by* or *following* those initial Twitter users.

### Step 2. Crowdsourcing identification of accounts

From the pool of potentially relevant Twitter accounts, we next seek to determine how they may be related to IBM. This task is crowdsourced by allowing users to associate one of the following types with each account:

- IBMer – someone currently working at IBM
- IBM "Group" – Twitter account related to an IBM business unit, office location, product, etc.
- Ex-IBMer – previously employed by IBM
- Non-IBM Affiliation

In the case of an IBMer type, users can choose a matching employee.

### Step 3. Email notification

Once a Twitter account is matched to a specific IBM employee, that employee is sent a notification email. To address privacy concerns, the employee is told if they take no action, the match would not be shown on the directory and tweets would not be collected.

### Step 4: Opt-in process

The employee could log in and view a "Twitter accounts" page of all their associated accounts. The user could take 1

[4] https://dev.twitter.com/docs/api/1/get/users/search

of 3 actions for a given Twitter account (*Figure **1***): "Share" it (listed in directory, tweets collected), "Private" it (further suppress even the Twitter account from appearing on the site, such as in searches), or mark it as "Not me" to report the incorrect identification.

**Step 5. Directory & Tweets for Identified Accounts**

Once the owner approves his/her Twitter account, tweets of the account are listed in the main column of the app in *Figure **2***. Each tweet has both Twitter account information and corporate identity at left/right sides. In addition, the stream of tweets can be filtered to Twitter accounts of only a specified type (like IBM groups), countries, tags and affiliations. Those filtering options show aggregated number of tweets and the top tags of all the tweets currently listed as well.

## Crowdsourcing Identification Task

Everyone who visits IBMersWhoTweet is regarded as a potential worker. We designed the task and feedback UI to motivate them to contribute. Each time a new page is loaded, the Help Find IBMers panel (bottom right of *Figure **2***), shows a Twitter account that is randomly selected from the pool of potentially relevant accounts, and gives options to clarify the type and identity of the account or skip to next person. The workers can choose either the type of the account (IBM group, ex-IBMer, Non-IBMer) or a matching employee from the corporate database. The system automatically displays the 3 most relevant employees from searching the corporate directory for the Twitter name. The worker can select any of these as a match. In cases when the initial automated search was not successful, the worker can also search with different keywords by using the "Search in IBM directory" input box.

When designing IBMersWhoTweet, crowdsourcing was chosen as the mechanism for identifying Twitter accounts over other automated methods such as machine learning or heuristic rules. Our early attempts with heuristic rules produced many false positives. We also found that it was often times trivially easy for a human to take multiple pieces of partially specified information to piece together the matching owning employee of a Twitter account. Employees also tended to use the same photo internally & externally making visual inspection against possible top matches easy. However, even automatic image recognition between internal & external photos of known matches failed to produce accuracy higher than 20%, due to subtleties in cropping, coloring, etc. As our attempts at automated approaches failed, we instead redirected our efforts into investigating mechanisms of increasing the quantity and quality of work produced by crowdsourcing the identification task.

Further, even if sophisticated algorithms could identify employees with high accuracy, crowdsourcing may positively effect the identified employee's willingness to join. Telling those employees that their "Colleagues found you." sets a much different tone than "Our algorithm found you."

As previously stated, the goal of IBMersWhoTweet is to aggregate tweets from IBM employees. In order to do this, it is critical that our workers contribute to identifying IBMers using Twitter. IBMersWhoTweet employs no financial rewards for these workers, but instead relies on peer competition, social reputation and altruism of visitors who come to use the web site. The feedback technique presented to the user is based on the crowdsourcing identification tasks they have completed and is shown in the "Contribution" section above the latest task they are presented (*Figure **2***, Top Right).

# Experimental Design

We made 7 different designs of achievement feedback to compare their motivational effectiveness (see Table 1). Each design implements a combination of three motivational techniques (Individual Achievement, Social Achievement, and Gamification). When a new user signs up, one of the experimental groups is randomly assigned to the user. Notice (in Figure 3) that Individual and Social achievements are separate UI components, thus stackable.

*Table 1. Combinations different motivational techniques*

| | | No feedback | Individual Achievement | Social Achievement | Both |
|---|---|---|---|---|---|
| Gamification | Off | *None* **(baseline)** | *Ind.* | *soc.* | *both* |
| | On | | *iG* | *sG* | *bothG* |

**Individual Achievement**

Individual Achievement is the number of Twitter accounts identified by the worker, which is grouped by account type. Below the numbers, thumbnail images of employees identified by the user are listed. (see the left column of Figure 3). The gamified version of individual achievement has additional UI components including the level badge and the number of employees to identify to reach the next level. Individual feedback is related to many motivational factors such as playfulness and self-development.

**Social Achievement**

Showing the worker's contribution as part of social activity is a common way to utilize *social capital* as a motivational source. The social achievement (Figure 3) shows how many IBM employees the site's workers were able to cumulatively identify (for various time-ranges). The gamified version of social achievements also includes a leaderboard-style ranking. To balance this mechanism in the non-gamified version, we included a list of recent contributors to the site.

## Hypothesis

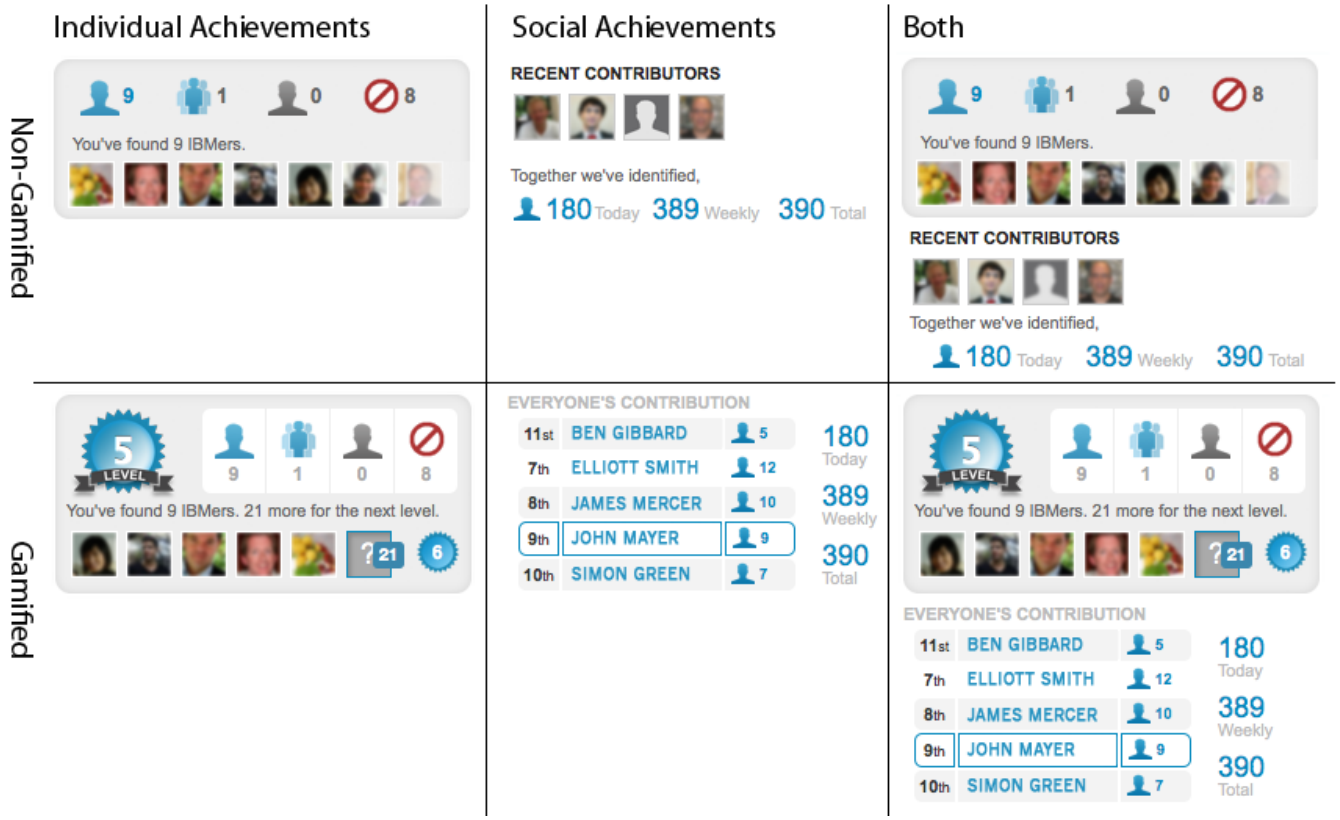


*Figure 3. Designs of achievement feedback UIs. "Individual achievements" (left) highlight the number of IBM employees found by the worker. "Social achievements" (middle) encourages the worker to think his/her contribution as part of group activity. "Both" (right) settings combine Individual and Social achievements. Gamification adds common playful items such as level badge, number of tasks till the next level and leaderboard. Profile photos and names are anonymized in the screenshots.*

We formulated hypotheses about the combinatorial effect of three motivational techniques (individual/social achievement, and gamification) on the quantity and the quality of identification tasks done by crowdsourced workers.

**H1.** Workers with no feedback will contribute the least.
**H2.** Providing more motivational elements will increase worker's motivation.
**H3.** Gamification will increase the amount of contribution.
**H4.** Gamification will lower the quality of contributions by encouraging workers to cheat.

The feedback UIs were tested as part of the IBMersWhoTweet system on August 6th 2012 and data described in this paper was collected thru February 22nd 2013. We bootstrapped the system by sending invitation emails to 499 users who had previously been identified as Twitter account owners in an earlier version of the system, but had not previously received an email invitation to join the site. 229 (45.9%) of those invited people visited the web site. Prior to release, a pool of 7,016 potentially relevant Twitter accounts was amassed using the methods described in Step 1 of *Figure 1*, which became the tasks assigned to workers who logged into the site. In addition to the invited users, an article was posted on a portal within the company to promote IBMersWhoTweet. Aside from those two mechanisms, the site grew organically by invitations to identified employees and through word-of-mouth. By February 22nd, an additional 2,917 users had logged into the site beyond those initially invited at the start of the experiment.

## Experimental Result

During the 6-month period, 3,144 users signed up for IBMersWhoTweet, and 437 (7.8%) of them completed at least one task. In total 4,154 Twitter accounts were classified: 2,427 (58%) IBM employees, 368 (9%) IBM groups, 225 (5%) ex-employees, and 1134 (27%) non-IBMers. Table 2 shows a detailed breakdown of tasks completed and identification type selected by each experimental group. We found that the behavior of workers under different conditions differs significantly according to a Kruskal–Wallis ANOVA test (chi-square=21.2, df=6, p=0.001) (Kruskal, and Wallis 1952). The *gamified social* design motivated workers to complete 10 times as many tasks as the *non-gamified social* did (U=1638, P<0.05)[5]. Overall, applying the gamification significantly increased the total number of completed tasks (U=27246, P<0.005).

[5] Mann-Whitney-Wilcoxon (MWW) rank-sum test is used to compare two non-parametric distributions.

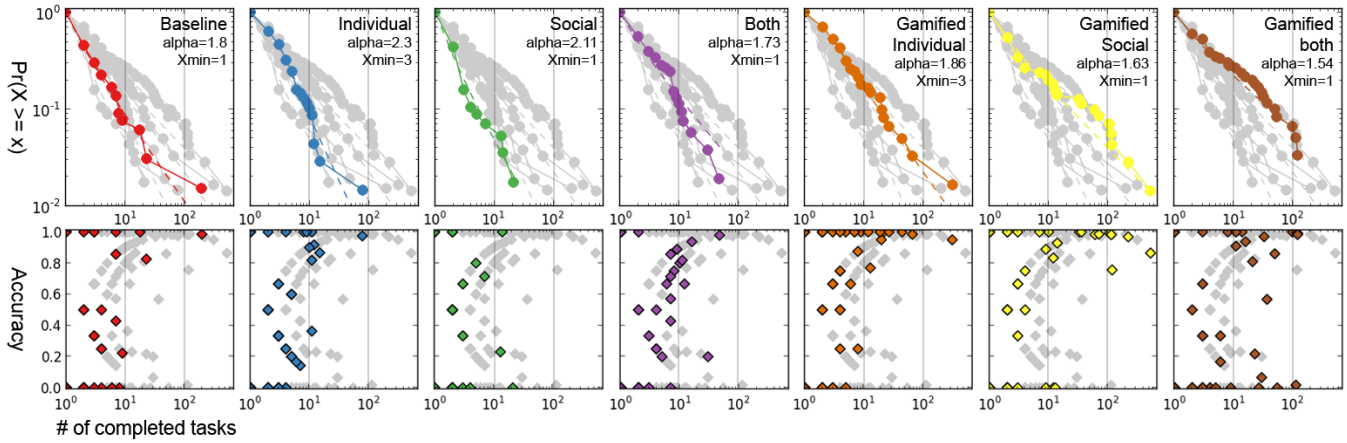


*Figure 4. Power-law distribution of the number of tasks completed (above) and the accuracy of them (below). In the Task Distribution graphs (above) the x-axis represents the number of tasks done in log scale, the y-axis represents the probability in log scale that a worker would end up finishing X tasks, where X is equal or greater than x-coordinate of the marker. Simply speaking, settings with higher curves are likely to get more tasks done. In the Accuracy graphs (below), the y-axis represents the accuracy of the tasks done by the worker. In general, workers who completed more than 10 tasks tend to be more accurate, with one exception (Gamified both).*

Those in the Gamified groups completed the majority (3062, 73.7%) of the entire tasks done, which supports H3 – *Gamification will increase the amount of contribution*.

To our surprise, the baseline ("None" column in Table 2) setting, which had no feedback UI, outperformed all the non-gamified designs (U=5395, P=0.27). This finding indicates that the task UI already provides strong intrinsic motivation, and additional feedback does not always yield more participation. Thus H1 (*Workers with no feedback will contribute the least*) does not hold.

The combinations of *individual* and *social achievements* (*Gamified/Non-gamified Both* settings) turned out not to be as effective as choosing a single better feature. This rejects H2 - *Providing more motivational elements will increase worker's motivation.*

We also found a significant interaction between *gamification* and *individual / social achievement* feedback. Among *non-gamified* feedback, *individual* feedback performed the best, while *social* feedback was the most effective among gamified designs.

Workers' behavior also followed Nielson's Participation Inequality rule (Nielson 2013) (also known as power law or Pareto principle), which states, "90% of users are lurkers who never contribute, 9% of users contribute a little, and 1% of users account for almost all the action." Due to the large variability over workers, standard charts and parametric statistical tests are not effective for comparing workers' behavior on different settings. Thus, we employed a log-log chart where both the x-axis (# of completed tasks) and y-axis (probability a worker would complete *x* number of tasks) are on log-scales, making power-law distributions almost straight lines (Clauset, Shalizi, and Newman 2009). The result (top row of Figure 4) allows at-a-glance comparison between experimental conditions. The higher a curve is, the more likely it was that workers would complete more tasks. For instance, one can easily see that the non-gamified Social (3rd from the left in Figure 4) is the least effective feedback UI. Some curves (*Baseline, Individual, Gamified Individual, and Gamified Social*) have outliers on their right end, which indicate a few enthusiastic workers who contributed most. Specifically with the *gamified social* design, it appears that the leader board mechanism, which is a combination of gamification and social feedback, seems to be very effective at motivating certain users to contribute a lot.

**Table 2. Numbers of Completed Tasks Per Worker With Different Feedback UI**

| | None | Non-Gamified | | | Gamified | | |
|---|---|---|---|---|---|---|---|
| | | Ind. | Soc. | Both. | Ind. | Soc. | Both. |
| # of users signed up | 424 | 465 | 456 | 432 | 458 | 486 | 423 |
| # of workers completed ≥1 tasks | 66 *15.6%* | 69 *14.8%* | 57 *12.5%* | 53 *12.3%* | 61 *13.3%* | 71 *14.6%* | 61 *14.2%* |
| **Tasks done per worker** | **6.00** | **4.46** | **2.40** | **4.74** | **11.59** | **20.69** | **14.52** |
| Employee | 4.56 | 2.13 | 0.68 | 3.34 | 6.56 | 11.35 | 9.13 |
| Group | 0.61 | 0.78 | 1.02 | 0.60 | 0.95 | 0.86 | 1.07 |
| Ex | 0.35 | 0.29 | 0.28 | 0.45 | 0.26 | 1.39 | 0.44 |
| None | 0.48 | 1.26 | 0.42 | 0.34 | 3.82 | 7.08 | 3.89 |

## Task Accuracy

In order to see how the experimental conditions effected the quality of crowdsourced work, we went through all the identified Twitter accounts one-by-one, and marked them as correct / incorrect[6]. The results in Table 3 show that task accuracy varies widely by feedback group ( $\chi^2 = 449, df = 6, p = 2.2e-16$ ). For example, *gamified both* feedback has the lowest accuracy of all those with *gamified* settings. Among the *non-gamified* groups, *social achievement* yields especially inaccurate results. Overall, tasks done by those with *gamified* feedback UIs are more accurate ( $\chi^2 = 164, df = 1, p = 2.2e-16$ ), which rejects H4

[6] For a small percentage (5%) of tasks (209 out of 4154) we were unable to determine correctness. Thus we conservatively counted them as incorrect, leaving the accuracies presented as lower bounds.

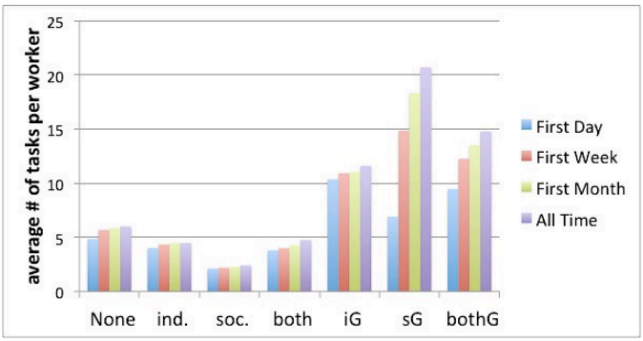


*Figure 5. Longitudinal effect of feedback UIs on task completion. While iG (Gamified Individual) is the most effective on the first day, sG (Gamified Social) has the strongest impact over time.*

- *Gamification will lower the accuracy by encouraging workers to cheat*.

**Table 3. Accuracy table of completed tasks**

| # of Tasks | | Non-Gamified | | | Gamified | | |
|---|---|---|---|---|---|---|---|
| | None | Ind. | Soc. | Both. | Ind. | Soc. | Both. |
| Completed | 396 | 308 | 137 | 251 | 707 | 1469 | 886 |
| Correct | 295 | 196 | 44 | 147 | 626 | 1260 | 535 |
| Incorrect or Undetermined | 101 | 112 | 93 | 104 | 81 | 209 | 351 |
| **Accuracy** | **0.75** | **0.64** | **0.32** | **0.59** | **0.89** | **0.86** | **0.60** |
| | | **0.56** | | | **0.79** | | |

For more detail, we plotted every worker according to the number of tasks he/she completed in log scale (X-position) and the worker's accuracy (Y-position), shown in the bottom row of Figure 4. The graphs show a general tendency that workers who completed more than 10 tasks are much more accurate (84%) than the rest (39%). While this tendency appears to be consistent over conditions, workers in the *gamified both* condition tend to be exceptionally inaccurate even when they contributed more than 10 tasks. Based on the overall population's accuracy, which supports our own experience that identification tasks are fairly easy to complete, those workers who submitted a large amount of incorrect results were likely putting very little effort into the tasks. We hypothesize that they enjoyed only the gamification features without paying attention to the task itself. However, any further explanation would demand additional surveys or interviews.

## Longitudinal Motivational Impact

During the 6-month experiment, new users consistently visited and contributed work to the site. At the same time, other workers were no longer completing tasks or even stopped visiting the site. Again, as completing tasks is critical for the overall value of the site (and thus incentive for anyone to visit it), maintaining and even growing the productive worker base is important. Thus we analyzed the longitudinal effects the conditions had on each worker's motivation for visiting the site and completing tasks.

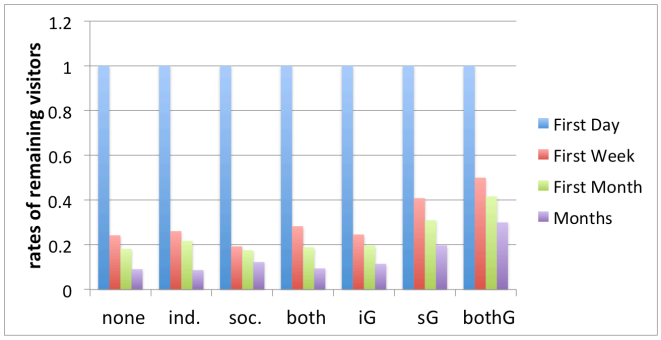


*Figure 6. Retention rates. bothG (Gamified Both) provides the strongest motivation for visiting the web site over time.*

Figure 5 illustrates the growth of tasks completed by workers. It is clearly visible that most conditions (five leftmost columns) had little success in motivating workers to contribute after the first day they joined. However, workers under *Gamified Social* and *Gamified Both* feedback continued to contribute over time. We hypothesize that the leaderboard element consistently motivated workers to contribute over time. It is also interesting to note that *Gamified Individual* and *Gamified Both*, which both have the *level badge*, were more successful in the first day than those in the *Gamified Social* group.

Figure 6 illustrates retention rates, which are the portion of users who keep visiting after the first day, week, and month. The graph clearly shows that retention rates drop off no matter what condition is assigned. However, that decline in return rate is not as sharp for both the *Gamified Both* and *Gamified Social* conditions as it was for the others. In fact, the number of users returning in the *Gamified Both* group after the first month was larger than the number of users returning in any of the other 5 groups after the first day. Interestingly, *the Gamified Individual* group decline in returning visitors was very similar to the other feedback groups. It appears that the leaderboard, in particular, rather than any type of gamification element, was effective at pulling users back to the site. Likely, this is due to a desire to check how their ranking/status has changed, however follow-up interviews are required to confirm this.

## Discussion

The results from our longitudinal experiment provide valuable insights into the design of crowdsourcing platforms:

• **Gamification is effective and does not necessarily cause cheating.** Our findings consistently indicate that quantity and quality of work, and longitudinal effect are all increased through gamification. However, it does appear that there was a certain threshold of gamified motivational feedback, above which, workers seem to lose their intrinsic motivation to complete a task accurately and instead work only to receive more achievement feedback and game the

system. This behavior is common in various incentive systems, such as GWAPs (von Ahn 2006) and beehive (Farzan et al. 2008). Finding the maximum safe level of gamification is crucial for motivational design. In our case, it was applying either of two gamified feedback elements (points/levels or a leaderboard) but not both.

**• The total motivational effect does not equal the sum of individual effects.** A common misconception about motivational design is that adding more features always increases the motivation. We consistently found that mixing weak motivational elements with stronger ones lowered the performance of the entire design. Sometimes, even 'no feedback' performed better than poor motivations, because the poor motivations lowered the intrinsic motivation the task already possessed.

**• Consider supporting short-term and long-term motivations with adaptive motivations.** The workers with the *gamified individual* condition performed the best for short-term periods, while the *gamified social* condition had the best long-term results. From this observation, we hypothesize an optimal design for a motivational feedback UI. First, this design would present gamified individual feedback, which would enhance users' short-term motivation (such as points / levels). After a certain period of time or behavior, the feedback would change to a mechanism that enhances users' long-term motivations (such as gamified social achievement elements, like a ranking/leaderboard). We see parallels in this approach with the design of online multiplayer games. Those typically start from single player mode (*gamified individual*) and gradually invite players to tournament mode (*gamified social*). But whether an adaptive approach will actually perform best, both in the short-term and long-term, without additional downsides such as encouraging workers to cheat, is beyond the scope of this paper, but worth further study.

There are several opportunities for future research. First, the study in this paper is based only on quantitative data. A qualitative study can provide additional insights and further explain the motivational structure of workers (e.g. why some workers cheated). For example, we anecdotally heard from one cheater that she did not pay attention to the task UI at all. Second, we conducted our experiment using the IBMersWhoTweet web site. Our site has some unique characteristics that might have effected workers' behavior in one way or the other. For example, they may have been more willing to help complete tasks on an internal company site without any motivational feedback than they would have been on a non-company affiliated site, such as Mechanical Turk. Running similar experiments with various types of tasks in different situations would be helpful to generalize the findings in this paper. Third, extending our approach to involve multiple motivational features, such as *triggers* or *incentives*, would be meaningful. Finally, another dimension we are interested in studying in the context of IBMersWhoTweet is the effect the relevance of a task to the worker has, in terms of quantity/quality of work completed. For example, a recommender system could be leveraged to personalize the tasks shown to workers by their interest (e.g. showing only people who work in similar areas as the user, or people who are geographically close).

## Conclusion

We explored the design space of achievement feedback UIs and tested each design in terms of their longitudinal effect on workers' motivation by comparing quantity and quality of tasks completed. Our experiment was conducted in the context of a novel web site that crowdsources the identification of Twitter accounts of employees of a company. As an internal enterprise service, financial rewards could not be leveraged to spur worker participation in completing tasks. We thus had to rely on other motivational mechanisms studied in this paper.

Overall, we found that *gamification* is an effective technique, but adding more motivational elements does not guarantee better performance. We also observed that short-term and long-term motivations are strongly affected by different feedback mechanisms.

The findings in this study complement our research understanding of contemporary gamification approaches (Cosley et al. 2003; Deterding et al. 2011; Shaw et al. 2011; von Ahn, and Dabbish 2008) and their potential value in designing effective crowdsourcing applications that are not driven by financial rewards.

## Acknowledgements

We are grateful to all the IBM employees who supported and/or participated in our experiment.

## References

Brabham, D.C. 2008. Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First Monday*, 13, 6, 1-22.

Brabham, D.C. 2010. Moving the crowd at Threadless: Motivations for participation in a crowdsourcing application. *Information, Communication & Society*, 13, 8, 1122-1145.

Clauset, A., Shalizi, C.R. and Newman, M.E.J. 2009. Power-Law Distributions in Empirical Data. *SIAM Rev. 51, 4* (November 2009), 661-703.

Chandler, D. and Kapelner, A. 2010. Breaking monotony with meaning: Motivation in crowdsourcing markets. University of Chicago mimeo

Chandler. D. and Horton, J.J. 2011. Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices. In *proceeding of: Human Computation, Papers from the 2011 AAAI Workshop*, San Francisco, California, USA, August 8, 2011

Cosley, D., Lam, S.K., Albert, I., Konstan, J., & Riedl, J. (2003). Is Seeing Believing? How Recommender Systems Influence

Users' Opinions. In *Proceedings of CHI 2003 Conference on Human Factors in Computing Systems*, Fort Lauderdale, FL, pp. 585-592.

Deci, E. L., and Ryan, R. M. 1985. *Intrinsic motivation and self-determination in human behavior*, Springer

Deterding, S., Dixon, D., Khaled, R. and Nacke. L. 2011. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (MindTrek '11). ACM, New York, NY, USA, 9-15.

Dow, S., Kulkarni, A., Klemmer, S. and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW '12). ACM, New York, NY, USA, 1013-1022.

Fogg, B.J. 2009. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology* (Persuasive '09). ACM, New York, NY, USA

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

Gibbons, R. 1996. *Incentives and Careers in Organizations. Incentive*, No. 5705, 1-37. Cambridge University Press.

Hackman, J., and Oldham, G. R. 1980. *Work redesign*, Addison-Wesley, Reading Mass.

Heyman, J. and Ariely, D. 2004. Effort for Payment: A Tale of Two Markets. *Psychological Science, 15*, (11, 2004), 787-793.

Howe, J. 2006. The Rise of Crowdsourcing, *Wired*, 14, 6.

Ipeirotis, P.G. 2010. Analyzing the Amazon Mechanical Turk marketplace. XRDS 17, 2 (December 2010), 16-21.

Kaufmann, N., Schulze, T. and Veit, D. 2011. More than fun and money . Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*. (Aug. 2011)

Kinnaird, P., Dabbish, L. and Kiesler, S. 2012. Workflow transparency in a microtask marketplace. In *Proceedings of the 17th ACM international conference on Supporting group work* (GROUP '12). ACM, New York, NY, USA, 281-284

Kruskal, W.H. and Wallis, W.A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47 (260): 583–621

Leimeister, J., Huber, M., Bretschneider, U., and Krcmar, H. 2009. Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition, *Journal of Management Information Systems*, 26, (2009) 197–224

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), ACM, New York, NY, USA, 77-85.

Nielson, J. Participation Inequality: Encouraging More Users to Contribute. Retrieved Feb 12, 2013, from Nielsen Normal Group. http://www.nngroup.com/articles/participation-inequality/

Preece, J. and Shneiderman, B. 2009. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction* (1) 1, pp. 13-32

Quinn, A.J. and Bederson, B.B. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 1403-1412.

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 *Proc. ICWSM11*, Barcelona, Spain.

Shaw, A.D., Horton, J.J., and Chen, D.L. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (CSCW '11). ACM, New York, NY, USA, 275-284.

Farzan, R., DiMicco, J., Millen, D. R, Brownholtz, B., Dugan, C., and Geyer, W., R. When the experiment is over: Deploying an incentive system to all the users. Persuasive Technology 2008.

von Ahn, L. Games with a purpose. 2006. *IEEE Computer Magazine,* 39(6):96-98, 2006

von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (Vienna, April 24-29). ACM, New York, 319-326, 2004.

von Ahn, L. and Dabbish, L. 2008. General Techniques for Designing Games with a Purpose. *Communications of the ACM,* August 2008. pp 58-67.