An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets

Jakob Rogstadius[a], Vassilis Kostakos[a], Aniket Kittur[b], Boris Smus[a], Jim Laredo[c], Maja Vukovic[c]

[a] Madeira Interactive Technologies Institute
University of Madeira
9000390 Funchal, Portugal
{jakob,vk}@m-iti.org

[b] Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
nkittur@cs.cmu.edu

[c] IBM T.J. Watson Research Center
Hawthorne NY 10532, USA
{laredoj,maja}@us.ibm.com

Abstract

Crowdsourced labor markets represent a powerful new paradigm for accomplishing work. Understanding the motivating factors that lead to high quality work could have significant benefits. However, researchers have so far found that motivating factors such as increased monetary reward generally increase workers' willingness to accept a task or the speed at which a task is completed, but do not improve the quality of the work. We hypothesize that factors that increase the intrinsic motivation of a task – such as framing a task as helping others – may succeed in improving output quality where extrinsic motivators such as increased pay do not. In this paper we present an experiment testing this hypothesis along with a novel experimental design that enables controlled experimentation with intrinsic and extrinsic motivators in Amazon's Mechanical Turk, a popular crowdsourcing task market. Results suggest that intrinsic motivation can indeed improve the quality of workers' output, confirming our hypothesis. Furthermore, we find a synergistic interaction between intrinsic and extrinsic motivators that runs contrary to previous literature suggesting "crowding out" effects. Our results have significant practical and theoretical implications for crowd work.

Introduction

This paper presents a study that assesses the effect of extrinsic and intrinsic motivators on task performance in the context of crowdsourcing markets. Crowdsourcing is a powerful approach to handling problems that by nature are difficult to solve computationally. The method is analogous to parallelizing computational work in programming environments and typically consists of segmenting the work into multiple small and independent pieces, which are then dispatched along with instructions through a crowdsourcing system to be solved by humans. Especially interesting forms of crowdsourcing are general-purpose task markets such as Amazon's Mechanical Turk (MTurk), in which a variety of different tasks can be posted. Popular crowdsourcing tasks include image tagging and classification, audio transcribing and various types of surveys. In return, the people who carry out the work are paid money for each completed task, often in small amounts: tagging an image, for example, may pay a few cents.

Crowdsourcing work involves a number of challenges different from those faced in traditional work settings. Crowd workers in general purpose markets like MTurk may have highly varying expertise, skills, and motivations. Employers ("requesters" in MTurk) have very little visibility into these characteristics, especially compared to a traditional organization in which workers are vetted during recruitment, have work histories, have reputations within and outside the organization, and may go through organizational socialization methods such as training to ensure they can appropriately satisfy their job requirements. Furthermore, workers can easily return work for a given job with no repercussions or even create an entirely new profile with a clear reputation. These challenges mean that employers have more limited means of eliciting high quality output than in traditional organizations.

This study experimentally assesses the interaction of extrinsic and intrinsic motivators in crowdsourcing markets using a novel experimental methodology that controls for self-selection effects and a novel experimental task that allows for a wide range of participant accuracy. For extrinsic financial rewards our results replicate previous studies suggesting that paying more does not result in more accurate performance. However, the presence of an intrinsic motivator did lead to more accurate worker performance. Furthermore, the interaction between intrinsic and extrinsic motivators appears to be such that workers provide highest quality results when intrinsic motivation is stronger than extrinsic motivation. Once extrinsic motivation takes over, accuracy converges to equal (and lower) levels regardless of the level of extrinsic motivation provided.

Related work

A traditional "rational" economic approach to eliciting higher quality work is to increase extrinsic motivation, i.e., how much an employer pays for the completion of a task (Gibbons 1997). Some evidence from traditional labor 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 & Palia (1995) found correlations between executive pay and firm performance when markets were allowed to self-regulate.

However, there is also evidence that in certain situations financial incentives may not help, or may even hurt. Such extrinsic motivations may clash with intrinsic motivations such as a workers' desire to perform the task for its own sake. For example, a classic experiment by Deci (1975) found a "crowding out" effect of external motivation such that 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 & Rustichini 2000). Given that crowdsourcing markets such as Mechanical Turk tend to pay very little money and involve relatively low wages (Ipeirotis 2010), external motivations such as increased pay may have less effect than requesters may desire. Indeed, research examining the link between financial incentives and performance in Mechanical Turk has generally found a lack of increased quality in worker output (Mason 2009)[1]. Although paying more can get work done faster, it has not been shown to get work done better.

Another approach to getting work done better 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 supported this hypothesis. For example, while crowdsourcing tasks which are framed in a meaningful context motivate individuals to do more, they are no more accurate (Chandler 2010). In summary, no approach has yet found extrinsic or intrinsic motivations to increase the quality of crowd workers' output[2].

However, there are a number of 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. Second, to our knowledge no study has yet looked at the interaction between intrinsic and extrinsic motivations; Mason & Watts (2009) vary financial reward (extrinsic), while Chandler & Kapelner (2010) vary meaningfulness of context (intrinsic) in a fixed diminishing financial reward structure. Finally, the task used in Chandler & Kapelner (2010) resulted in very high performance levels, suggesting a possible ceiling effect on the influence of intrinsic motivation.

Crowdsourcing and Mechanical Turk

Amazon's Mechanical Turk (MTurk) is a general marketplace for crowdsourcing where requesters can create Human Intelligence Tasks (HITs) to be completed by workers. Typical tasks include labeling objects in an image, transcribing audio, or judging the relevance of a search result, with each task normally pay a few cents (USD).

Work such as image labeling can be set up in the form of HIT groups, where the task remains identical but the input data on which the work is carried out varies. MTurk provides a logical workflow within such groups where workers are continuously offered new HITs of the same type after they accept and complete a HIT within the group. MTurk also allows splitting a HIT into multiple identical assignments, each which must be taken by a different worker, to facilitate for instance voting or averaging schemes where multiple workers carry out the same task and the answers are aggregated.

Running Controlled Studies on MTurk

Using MTurk poses a problem for experimental studies, since it lacks support for random participant assignment, leading to issues even with between subjects control. This is especially problematic for studies of motivation, as self-selection is an inherent aspect of a task market. This means that results in different conditions could be due to attracting different kinds of people rather than differences in the conditions themselves. In this study, given two tasks of which one pays more and one pays less, making both of them available on the site at the same time would bias the results (contrast effect)[3]. If they were put up at different times, then different workers might be attracted (e.g., Indian workers work at different times than Americans; some days/times get more activity than others, etc.), or more attractive work could be posted by another requester during one of the conditions but not the other..

The other extreme is to host everything on the experiment server, using MTurk only as a recruitment and fulfillment host. All participants see and accept the same identical task, and are then routed to the different places according to the appropriate condition on the experimenter's side. This fails when studying how workers act naturalistically, as everything is on the host environment. Thus aspects such as the title, description, and most importantly reward cannot be varied by condition, making it impossible to study natural task selection.

This study proposes a novel approach in which participants fill out a common qualification task with neutral title

[1] The relationship between price and quality has also had conflicting results in other crowdsourcing applications such as answer markets (e.g., Harper et al., 2008)

[2] Though there are other methods; for example, Kittur et al. (2008) used a variety of methods to increase signal in subjective tasks, such as signaling monitoring or increasing the cost of bad faith answers. Another example is CrowdFlower's "gold standard" approach, which provides feedback to workers when they answer specific sampled questions incorrectly. However, these are task-specific approaches that may not work for many kinds of tasks, and while they may filter out poor quality work by raising the threshold for acceptance, may not motivate high quality output.

[3] This contrast effect would be problematic even for non-simultaneous posting if workers saw one task at one price and then the same task at another price at a later time.

and description. This qualification task (in our case, simply collecting demographic data) is hosted on the experimenter's server and on completion randomly assigns the participant to one of the conditions through a condition-specific qualification in the MTurk system. This qualification enables workers to see and select only tasks in that condition when searching for tasks in the natural Mturk interface. In this study we used an MTurk qualification type with six different possible values corresponding to the different conditions. The key benefit of this approach is that participants still use the MTurk interface as they naturally do to self-select tasks, which can have condition-specific titles, descriptions, content, and rewards. While participants can still explicitly search for the tasks in other conditions and see them in some HIT listings, HITs cannot be previewed without having the appropriate qualification. Hosting the task externally (which we did not do) would avoid the explicit search problem, but would not address non-preview textual descriptions or the key issue of supporting condition-specific variations in payment.

Another advantage of the qualification-task-approach is that the worker will always retain the qualification granted to them by the experimenter (so they can be kept track of). Thus, for example if an experimenter wanted to make a new experiment available to a subset of their participants they could add the qualification for it to the appropriate participants and the task would automatically become available to the target participants on MTurk. For more intensive recruitment, once a worker has completed the qualification task and their worker ID is known, they can be emailed directly by the experimenter, even if they did not complete an experiment.

This proposed approach for recruiting participants from a crowdsourcing market lets us retain some of the control of a traditional laboratory setting, the validity of participants searching for work in their natural setting, and the benefits offered by a greater diversity of workers more representative of the online population than undergraduates would be (Horton et. al. 2010). The legitimacy of doing both cognitive and social experiments with Mechanical Turk has been supported by multiple studies, e.g. (Heer 2010; Ipeirotis 2010).

Study

With the goal of measuring the interaction effects of intrinsic and extrinsic motivation on Amazon's Mechanical Turk, we decided on a 2x3 design for our experiment. We operationalized our motivation manipulation through two levels of a "cover story" (non-profit, for-profit, described in more detail below) and three levels of reward (0, 3, and 10 cents USD). We then designed a task that allows us to quantitatively measure the quality of the work in a way where quality is dependent on effort while avoiding ceiling effects. Based on the results from previous work, we worked primarily with four experimental hypotheses:

H1 Tasks in the non-profit (i.e. charity) conditions will be completed faster than tasks in the for-profit conditions.

H2 Tasks in the non-profit (i.e. charity) conditions will be completed more accurately than tasks in the for-profit conditions.

H3 Tasks in high-pay conditions will be completed faster than tasks in low-pay conditions.

H4 Tasks in high-pay conditions will be completed more accurately than tasks in low-pay conditions.

Recruitment

To recruit participants, a Human Intelligence Task (HIT) was posted on Amazon's Mechanical Turk (MTurk), appearing to be from a fictitious organization that handles crowdsourcing on behalf of third party pharmaceutical and health-related organizations. The HIT advertised that by completing the associated questionnaire workers would obtain a qualification to complete further HITs. The HIT consisted of an externally hosted questionnaire that collected broad demographic data from participants, as well as data on their experience on MTurk. Once completed, the questionnaire allocated participants to one of the six experimental conditions by assigning them one of six different qualifications on MTurk, and in addition awarded participants a one-off bonus of 2 cents USD.

Upon completing the questionnaire and obtaining a qualification, participants gained access to further HITs in their assigned condition. These HITs could be accessed either through a link provided at the final confirmation page of the qualification form, through an email sent to them or through regular search.

A worker who would list all work currently available from the fictitious organization could at any given time see six HIT groups with generic and identical titles ("Medical image analysis") and descriptions ("See HIT preview for instructions") but with different payment levels and requiring different qualifications. However, workers listing work available to them would only see the HIT group relevant to their qualification (if any), and in any case could preview only the qualification-relevant HIT group to see a detailed description and image of the actual task.

On average, a little over a day elapsed between when working participants submitted the questionnaire and when they accepted the first task. However, there was a significant dropout effect in which most workers who went through the registration process (81.3%) did not complete any experimental tasks at all.

Experimental Task

The experimental task consisted of a single HTML page that included a cover story at the top of the page, instructions on how to complete the task, an image to analyze, and input fields for answers. The cover story for the non-profit condition was:

> The **Global Health Council**, a nonprofit organization and the world's largest membership alliance dedicated to saving lives by improving health throughout the

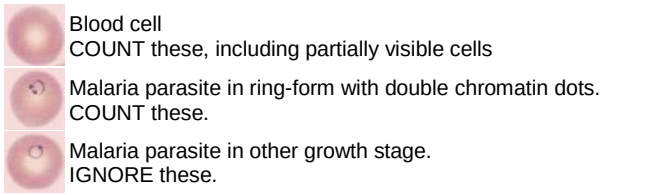

Figure 1. Instructions given to participants on how to complete the experimental task.

world, is running a study to assess the effectiveness of recent advances in the treatment of malaria.

The for-profit statement gave the same information, except that the organization was changed to "**Rimek International**, a major actor in private pharmaceutical manufacturing". The instructions for all conditions then had as follows:

> This task requires you to identify blood cells infected with malaria parasites. The malaria parasite goes through a number of growth stages. For this task you are required to identify the parasites that are in a specific growth stage (ring-form with two adjacent dots). Look at the image below and
>
> 1) Count the number of malaria parasites in ring-form, having double chromatin dots.
>
> 2) Count the total number of blood cells in the image.
>
> Some images may be ambiguous and require guesses or estimates. Please keep in mind that the quality of any such estimates will directly influence the quality of this research.

The instructions concluded with a legend of objects featured in the task image (Figure 1).

After the instructions, participants were shown a computer-generated image with known properties (Figure 2), and were asked to enter i) the number of malaria parasites in the correct growth stage in the image, and ii) the total number of blood cells.

The experimental images were generated by independently varying the number of cells in the image, and the number of the malaria parasites that participants had to count. Around 18% of cells contained noise in the form of parasites in non-interesting growth states and images with high cell counts had significant visual overlap of the cells. Initial tests showed that the experimental tasks would take between 30 and 200 seconds to complete, with an average of around one minute.

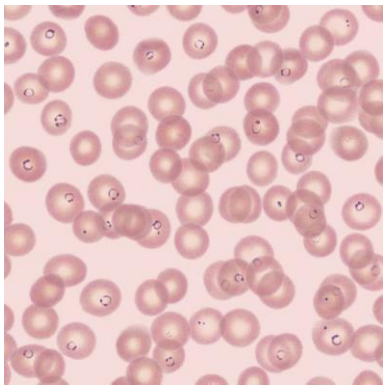

Figure 2. A sample image of medium complexity from the experimental task.

Upon completion of the HIT, participants were automatically given the option to complete further HITs in the condition. By clicking on the "accept" button, participants could attempt another HIT. Each condition consisted of 100 HITs, with two identical assignments per HIT.

Results

The study ran for 48 days. Once all the advertised HITs in a condition were completed no more participants were allocated to that condition. However, neither 0 cent conditions attracted enough workers for all tasks to be completed; an issue we will return to later. To minimize bias all conditions appeared to be available in the public listing, even though all work was completed for some conditions and they did not accept new participants. In a few instances (4.7%) assignment answers had been swapped for parasite and cell counts. These answers were manually corrected when the answers differed by more than 25% and when the error was lower after swapping.

The 3 and 10 cent rewards were based on an estimated average task completion time of one minute, which would have yielded hourly wages of $1.8 and $6 USD respectively. In practice, however, participants spent more time than estimated per task and achieved effective hourly wages of only $1.4 and $3.3 for the 3 and 10 cent groups.

Demographics

A total of 843 people completed the qualification questionnaire, of which 158 showed up, i.e. completed at least one assignment. Unless otherwise stated, these are the participants to which the results refer. Of the participants that showed up, 49% were female. In addition, 42% reported having lived only in South Asia (including China and India) and 35% only in North America (excluding Mexico). Participants from South Asia compared to those from North America on average had lower yearly income (median <$5k vs. $20k-$60k), higher education and were younger. The median working participant had a bachelor's degree and was 25-34 years old.

Metrics

For each experimental task (assignment) the following information was collected: reported cell count, reported parasite count, time spent and participant ID. In addition, for each participant the following information was recorded by the questionnaire: demographics (gender, age, education, income, region(s) of residence), time registered on MTurk, weekly time spent on MTurk, diseases affecting user or somebody close to them (including malaria), previous experience with blood analysis.

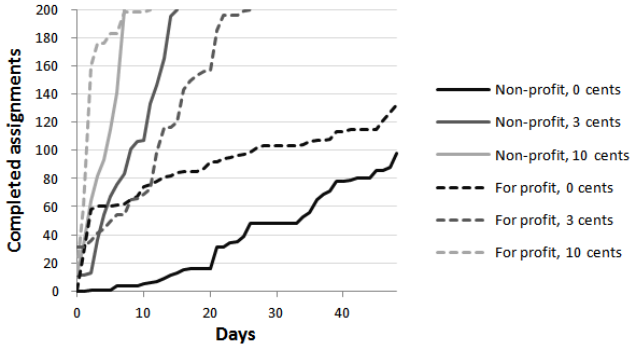

Figure 3. Time taken to complete each condition's batch of assignments. Contributions from individuals who chose to complete many assignments in a sequence show up in the graph as vertical jumps in the time series.

To measure the effort that each participant chose to spend, we use total completed assignments, total working hours and mean time per task. Uptake ratios (ratio of registering participants who completed at least one task) are also reported, as they have implications for total work completion rates.

An aggregate accuracy metric was defined as follows to capture quality of answers

$$accuracy = 1 - \frac{1}{2}\left(\frac{|p_{est} - p_{real}|}{p_{real}} + \frac{|c_{est} - c_{real}|}{c_{real}}\right),$$

where p is parasites and c is cells.

A combined metric for task complexity was also introduced, with greater weight given to parasites than to cells as participants had to consider the growth stage of the parasite when counting them:

$$complexity = c_{real} + 3p_{real}.$$

Work Effort

Figure 3 shows the rate at which the assignments in each condition were completed, with higher rates for higher paying conditions. The progress rate is not steady since most progress comes in short bursts from single individuals who choose to complete many assignments in one go.

Most participants chose to complete only a few tasks, and the distribution was heavily skewed (mean 6.5, median 2). Figure 4 shows how the total workload was distributed among participants. The graph shows that a single participant contributed half the total work in the for-profit 0-cent condition, and that as much work was produced by a single

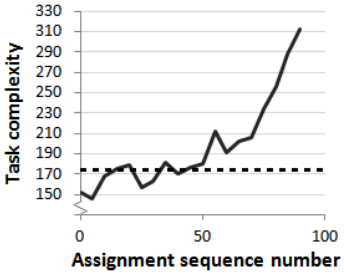

Figure 5. Average task complexity by assignment sequence number. The dotted line shows the average complexity in the entire workload.

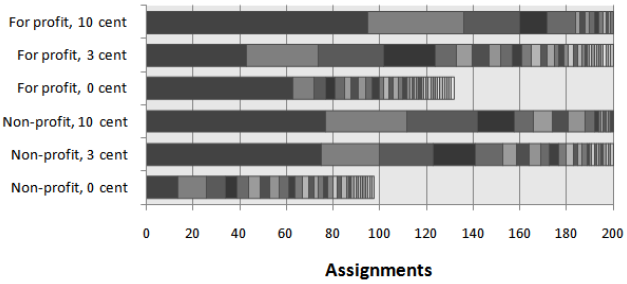

Figure 4. Distribution of completed assignments among participants. Each participant is represented by one bar segment. The two 0 cent workloads did not complete.

participant in the for-profit 10-cent condition as was produced in total in the non-profit 0-cent condition. This distribution can be expected when workers are allowed to self-select how many tasks to complete and it is representative of normal work distribution on MTurk.

Table 1 lists various indicators of interest, including uptake (percentage of registering participants who completed at least one assignment). Payment variations had clear effects, with total uptake numbers of 12.9% of registering participants in the 0-cent category, 25.1% in the 3-cent category, and 39.2% in the 10-cent category.

Further analysis of the data shows that the average task complexity for the first assignment completed by each participant was lower (158) than the average complexity among all tasks (173). As MTurk presents participants with tasks in random order, a significant deviation from batch average for the first task means that uptake is affected by the upfront complexity. Payment level affected this first-task complexity average with scores for the different payment groups being 139 for 0 cents, 169 for 3 cents and 181 for 10 cents. The for-profit group averaged at 153 and the non-profit group at 164. Figure 5 shows how task complexity changed as participants completed more tasks. The expected average task complexity was only achieved after participants had completed 15 assignments, while participants completing many HITs had a very high average complexity because they completed difficult HITs that others presumably chose not to work on.

Participants' region of residence also affected performance. As mentioned previously, 42% of participants

		Completed assignments	Mean complexity of completed assignments	Working participants	Uptake	N. American workers	S. Asian workers	Female workers	Assignments/worker	Mean task accuracy
Non-profit	0 cent	98	128,3	32	12%	28%	54%	52%	3,1	0,83
	3 cent	200	173,1	26	31%	35%	51%	56%	7,7	0,83
	10 cent	200	173,1	16	41%	31%	49%	41%	12,5	0,73
For profit	0 cent	132	147,6	36	14%	31%	53%	56%	3,7	0,71
	3 cent	200	173,1	33	22%	35%	47%	45%	6,1	0,66
	10 cent	200	173,1	15	38%	48%	35%	38%	13,3	0,75

Table 1. Performance metrics for the six conditions. Uptake refers to the ratio of qualified participants who chose to complete at least one assignment.

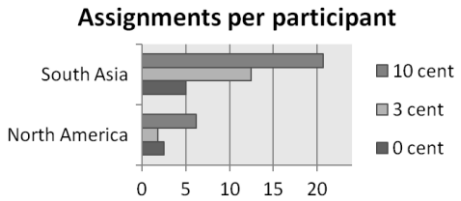

Figure 6. Breakdown of total work effort (average assignments per participant) by payment level and participant location.

reported having lived only in South Asia, while 35% had lived only in North America. Yet, 72% of assignments were completed by Asians and 15% by Americans. On average participants from North America completed 2.8 HITs with 89% accuracy and 123.5 mean task complexity, while those from South Asia completed 11.2 hits with 71% accuracy and 172.9 mean task complexity.

The effect that variations in payment had on the number of completed assignments per participant in these two worker groups can be seen in Figure 6. A two-way between-groups ANOVA showed a significant main effect of location [$F(1, 115)=10.9$, $p=0.001$] and payment [$F(2, 115)=3.5$, $p=0.034$]. The effect of both of these variables was moderate (eta squared=0.086 and 0.057 respectively). Post-hoc comparisons using the Tukey HSD test indicated significant differences only between means of the 0 and 10 cent groups. While increasing payment levels generally lead to increased work effort for participants both from South Asia and North America, going from a 0 to 3 cent reward appears to have had no effect on Americans.

We also observed that using a non-profit cover story slightly increased uptake and average assignments per participants for Americans and decreased it slightly for Asians, and while these observations are similar to results by Chandler & Kapelner (2010) the effects in our study were not statistically significant. Differences in both work effort and accuracy based on region of residence were clearer than differences based on income.

The time which participants spent on tasks of low complexity was consistent across conditions (Figure 7). It then began leveling off around complexity of 100, but peaked at different levels for different conditions. Participants in the for-profit and 10-cent groups spent more time working on complex tasks than participants in the other conditions. Working time decreased significantly for all conditions at the highest complexity levels.

Accuracy

The second metric of interest is work quality, which we quantify as *accuracy*. An ANOVA showed a significant main effect of "cover story" (non-profit, for-profit) on the accuracy of completed tasks ($F(1,1024)=38.1$, $p<0.0001$), while reward had no significant effect. Although we report accuracy scores here based on absolute errors, these errors were almost exclusively underestimates of the true values.

Returning to Figure 7, we see the effects on task accuracy and time spent on tasks from increasing levels of task complexity. Accuracy decreased with increasing complexity in all conditions, with participants doing non-profit work being consistently more accurate than for-profit workers. While there appears to be no correlation between time spent on a task and achieved accuracy, task times reported in MTurk are generally not reliable as workers often may have multiple windows and tasks open at once.

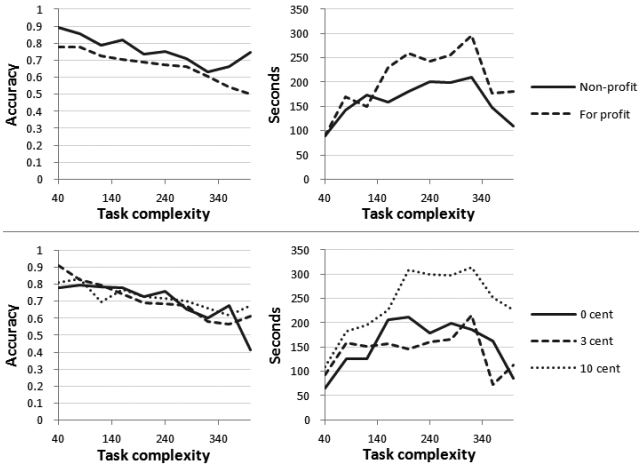

Figure 7. The effect of variations in task complexity on task accuracy (left) and time spent per task (right).

Table 1 also shows how the mean complexity of completed assignments in the two incomplete 0-cent conditions was lower than in the four completed conditions. This indicates that participants chose to complete only the easy tasks, presumably because the incentives were too small to motivate the effort of working on the most complex tasks. As accuracy decreased with increasing task complexity, this selection effect needs to be accounted for when comparing the mean accuracy between conditions and we thus conclude that participants in the 3-cent non-profit condition produced the most accurate results.

Finally, Figure 8 considers how accuracy changed as participants completed more tasks, suggesting that participants under both cover stories performed equally well in their first three assignments. After this, non-profit participants kept gaining in accuracy up to the seventh task, while for-profit participants became less accurate. Beyond this point up to the 25th assignment, both groups became increasingly less accurate, but the performance of non-profit workers decreased slightly slower than others'. The data showed no further accuracy decreases beyond the 25th task, but sample sizes for these levels were limited to only a handful of workers. The decrease in accuracy from increasing numbers of completed tasks cannot be explained by the associated increase in task complexity (Figure 5), as the average complexity change for the first 25 tasks was

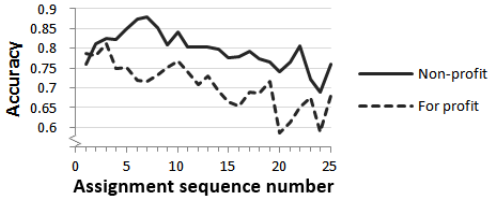

Figure 8. Mean assignment accuracy by asgmt. sequence number.

too small. As most participants completed only a few assignments, the number of samples on which the series are based decreases rapidly along the horizontal axis and the increasing variance seen in the graph is to be expected. The sample size was not considered large enough to present similar data across payment groups.

Discussion

Our motivation for the study was to experimentally assess how workers' performance and effort is affected by varying the levels of intrinsic and extrinsic motivation in a task, as well as examining interaction effects between the two motivational factors. To assess the work completed as part of this study, we measured completion speed and accuracy.

Consistent with prior work, we found that paying people more did not lead to increases in their output accuracy. However, unlike previous work we did find a significant effect of intrinsic motivation on output accuracy: people were more accurate under the non-profit framing than they were under the for-profit framing. Not only was this true for the average task, but also for assignment sequences (Figure 8) and for varying levels of task complexity (Figure 7). The intrinsic motivation frame did not impact uptake speed; specifically we saw no change in batch completion speed (Figure 3), completed tasks per worker and worker uptake (Table 1).

We also observed interaction effects between intrinsic and extrinsic motivation, resulting in changes in worker accuracy between conditions that cannot be explained by linear models (Table 1). One explanation of these findings consistent with prior theory (e.g., Deci, 1975) is that intrinsic motivation has a strong positive effect on worker accuracy, but only until the point where extrinsic factors become the main motivator. Further work is needed to explore this and other possibilities.

The hypothesis that increased payment increases work output is confirmed by the data, in full agreement with results from previous studies (Mason 2009). Higher rewards substantially increased both participant uptake and overall completion rates. This effect may be further strengthened by that MTurk design gives less exposure to low-paying HITs, as it is easy to sort available HITs by reward. Paid participants were also more tolerant to task complexity, as indicated by the average first-task complexities as well as the lower average complexities of completed tasks in the two 0-cent conditions. Participants in the for-profit 10 cent condition in fact exhibited higher-than-average task complexity for their first task. We also find it interesting that although progress in the 0-cent conditions was significantly slower than in the paying conditions, 12-14% of workers in a task market built around extrinsic motivation were still willing to contribute some work without any form of payment.

Figure 6 shows that participants from both South Asia and North America greatly increased their workload once sufficient payment was reached. We note, however, that this sufficient level appears to differ between regions and that Asians were willing to work for less compensation than Americans. The data in Figure 5 together with that regional differences were greater than differences between income groups, suggests that both of these rates were high enough to have an effect on Asian workers (from a lower-income society), while Americans (from a higher-income society) and others perceived the 3 cent reward as equal or worse than working without compensation. This finding is in agreement with previous studies showing that if the extrinsic motivation (in this case the reward) is not adequate, performance is likely to suffer (Finin, 2010).

Sample bias

Studies of motivation on MTurk, including ours, need to address problems introduced by large differences in sample size for different participants, such as a large number of tasks completed by a small group of participants. This distribution is natural to crowdsourcing markets (and many online communities) in which workers self-select which and how many work items to complete. As our goal is to measure effects of motivation on total work output, our analyses consider the task as our unit of analysis; however, we note that this assigns more weight to people who contribute more work.

An alternative would be to use the worker as the unit of analysis (e.g., calculate means for each worker, followed by taking the means of those means for each condition). In our study, this would have not only biased results towards workers who we know only completed one or two tasks each, but also introduced noise from the great variations in workers' mean task complexity, as well as not being representative of the natural distribution of task uptake.

Strategies and Guidelines for Crowdsourcing

Below we discuss guidelines suggested by our findings for crowdsourced work.

Speeding up Progress

The importance of adequate payment on a crowdsourcing market like MTurk is crucial. Not only did higher paying tasks attract workers at a higher rate; those workers also completed more work once they showed up. This resulted in both higher and more predictable rates of progress. The effect which payment has on progress is simple; higher payment leads to quicker results.

In addition to increased payment, the data shows that quicker results can be achieved by simplifying each work item, which in turn increases uptake of workers.

Our results show no effect of intrinsic motivation on work progress. However, uptake might be improved by highlighting intrinsic value in task captions and summaries, something we could not do due to our study design.

Increasing Accuracy

Emphasizing the importance of the work (in this case working for a non-profit organization) had a statistically significant and consistent positive effect on quality of answers in the study. Effects were particularly strong at lower payment levels, with differences in accuracy of 12% and 17% for the 0 and 3 cent conditions. These marked differ-

ences are surprising given the similarities between the conditions, which both included malaria and the only difference being the company the task was being done for. This difference between conditions was even more conservative than Chandler & Kapelner (2010), who either gave workers a description of purpose or did not.

The results may have application to crowdsourcing charity work, suggesting that lower payment levels may produce higher quality results. It is unlikely that workers actually prefer to work for less money, thus this might suggest that intrinsic value has to be kept larger than extrinsic value for the accuracy benefits to appear.

Although in this paper we specifically investigate the non-profit/for-profit distinction as our method of investigating intrinsic motivation, there are a number of other possible ways for affecting intrinsic motivation as well. Future work investigating factors such as social identity, goal setting, and feedback could all be profitable directions (Cosley et al., 2005; Ling et al., 2005).

Demographic Considerations

Although most work in this study was performed by participants from Asia, people from North America were on average more accurate but less tolerant to high task complexity. Such regional differences are worth keeping in mind for a number of reasons. Americans are a large group; a third of the workforce in our study and 40% of site visitors according to statistics from Alexa (www.alexa.com). In addition, nationality is one of the few built in ways of restricting access to work that MTurk supports, without creation of additional qualification tasks. Our data does however suggest that excluding Asian workers is likely to have severe impacts on work completion rates, in particular if payment is kept at a level which is perceived by Americans as low.

While 158 participants completed work in this study, only nine completed 30 or more assignments and together account for half the total output. Designing tasks that attract these workers may have significant effects on work completion rates and their demographics are therefore worth mentioning. All carried bachelor's degrees or higher and all but one lived in South Asia. Six were male and ages were equally distributed between 18 and 44. Most reported spending more than six hours per week working on MTurk and yearly incomes were generally below $5,000. The participants were equally distributed between the cover stories, but favored higher paying tasks. The highest work output (95 assignments) was by an Asian woman, 35-44 years old with a bachelor's degree and with a yearly income between $20,000 and $60,000.

Conclusion

This paper has shown that work accuracy can be improved significantly through intrinsic motivators, especially when extrinsic motivation is low. We also find, consistent with prior work, that increasing levels of payment increases work output regardless of intrinsic value. In addition to the findings, we present a qualification-based approach to conducting experimental studies on Amazon Mechanical Turk that addresses the need for both control and realism.

Acknowledgements

This work is supported in part by an IBM Open Collaboration Award; by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy); and by NSF grants OCI-0943148 and IIS-0968484.

References

Chandler, D. and Kapelner, A. May 2010. *Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets*, working paper, http://www.danachandler.com/files/Chandler_Kapelner_BreakingMonotonyWithMeaning.pdf.

Cosley, D et al. 2005. How Oversight Improves Member-Maintained Communities, In *Proc. CHI 2005*, 11-20. ACM Press.

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

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

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL Workshop on Creating Speech and Text Language Data With Amazon's Mechanical Turk*. Association for Computational Linguistics.

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

Gneezy, U., Rustichini, 2000. A. Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 791-810. MIT Press.

Harper, F. et al. 2008. Predictors of Answer Quality in Online Q&A Sites, In *Proc. CHI 2008*, 865-874. ACM Press.

Heer, J., Bostock, M. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proc. CHI 2010*, 203-212. ACM Press.

Horton, J., Rand, D., Zeckhauser, R. 2010. The online laboratory: Conducting experiments in a real labor market. NBER Working Paper w15691.

Hubbard, R. G., Palia, D. 1995. Executive pay and performance: Evidence from the US banking industry. *Journal of Financial Economics*, 39(1), 105–130.

Ipeirotis, P. 2010. Demographics of Mechanical Turk. New York University Working Paper.

Kittur, A., Chi, E., Suh, B. 2008. Crowdsourcing User Studies With Mechanical Turk. In *Proc. CHI 2008*, 453-456. ACM Press.

Lazear, E. 2000. Performance, Pay and Productivity. *American Economic Review*, 90 (5), 1346–1361.

Ling, K et al. 2005. Using Social Psychology to Motivate Contributions to Online Communities. *JCMC*, 10 (4).

Mason, W. A., Watts, D. J. 2009. Financial incentives and the performance of crowds. In *Proc. ACM SIGKDD Workshop on Human Computation*, pp. 77-85. ACM Press.