View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Harvard University - DASH

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

The Effects of Performance-Contingent Financial Incentives in Online Labor Markets

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation	Yin, Ming, Yiling Chen, and Yu-An Sun. In press. The effects of performance-contingent financial incentives in online labor markets. In The proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13), 14-18 July 2013, Bellevue, Washington, USA. Palo Alto, Calif: AAAI Press.
Accessed	February 19, 2015 1:10:22 PM EST
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:11129153
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

(Article begins on next page)

The Effects of Performance-Contingent Financial Incentives in Online Labor Markets*

Ming Yin
Harvard University
mingyin@fas.harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

Yu-An Sun
Xerox Innovation Group
YuAn.Sun@xerox.com

Abstract

Online labor markets such as Amazon Mechanical Turk (MTurk) have emerged as platforms that facilitate the allocation of productive effort across global economies. Many of these markets compensate workers with monetary payments. We study the effects of *performance-contingent* financial rewards on *work quality* and *worker effort* in MTurk via two experiments. We find that the magnitude of performance-contingent financial rewards alone affects neither quality nor effort. However, when workers working on two tasks of the same type in a sequence, the change in the magnitude of the reward over the two tasks affects both. In particular, both work quality and worker effort increase (alternatively decrease) as the reward increases (alternatively decreases) for the second task. This suggests the existence of the *anchoring effect* on workers' perception of incentives in MTurk and that this effect can be leveraged in workflow design to increase the effectiveness of financial incentives.

1 Introduction

Crowdsourcing has become a new form of production where a global population of workers make short-term contributions to tasks of their choice in online workplaces. For example, in Amazon Mechanical Turk (MTurk), an online labor market for micro-tasks, every day there are more than 50,000 – 240,000 new tasks arrived and similar number of existing tasks completed according to MTurk Tracker[1]; citizen science websites such as Zooniverse and eBird have attracted a large number of non-expert volunteers making contributions to scientific projects in a wide array of domains, including space, climate, nature, and health.

Crowdsourcing brings great opportunities to AI research. A fundamental problem of interest is how to integrate human and machine intelligence to improve productivity of crowdsourcing systems. AI techniques have been used in designing complex workflows and integrating worker contributions in online labor markets (Dai, Mausam, and Weld 2010; 2011; Lin, Mausam, and Weld 2012a; 2012b; Kamar, Hacker, and Horvitz 2012). Such work often assumes an inherent error rate for each worker and uses learning or decision-theoretic methods to estimate the ability of individual workers as well as the inherent difficulty of tasks to decide on when to ask more workers to contribute on a task.

However, human workers are motivated by various extrinsic and intrinsic motives, such as monetary rewards and the enjoyment of completing a task (Benkler 2002). They may be incentivized to exert more or less effort and be influenced by their psychological biases when the design of tasks or workflows affects these motives. Thus, a thorough understanding of how incentives affect worker effort and productivity in crowdsourcing is important for developing methods to improve the productivity of crowdsourcing systems.

In this paper, we experimentally study the effects of a particular type of incentives, *performance-contingent* financial rewards, in online labor markets where requesters post tasks with specified monetary compensation and workers choose which tasks to work on and receive payments for work completed. Unlike in traditional labor markets, the size of tasks as well as the amount of payments are often much smaller and workers have much higher mobility in online labor markets. These differences make the classical, fundamental question of the relationship between financial incentives and productivity relevant again for online labor markets. By "performance-contingent", we mean that the amount of the reward for a task depends on the quality of work produced, where the quality is evaluated according to some metric of interest to the task requester. Although not all crowdsourcing tasks use performance-contingent rewards (e.g. the quality of work for some tasks is subjective, not verifiable, or is too costly to be practical to verify), such rewards are commonplace across traditional labor markets, making them a good candidate to study for online labor markets.

A few recent studies explored the effects of financial incentives in MTurk from several perspectives. Mason and Watts (2009) examined performance-independent financial rewards in two experiments, where workers were paid a fixed reward for each task completed and had the option of continuing to work on more tasks. They found that workers chose to complete more tasks when the magnitude of the fixed reward increased, but the work quality was not improved. Rogstadius et al. (2011) made a similar observation in their experiments. Arguably, it is not surprising that the magnitude of fixed rewards does not affect the work quality — after all, a worker is better off completing a task with

*We thank the support of Xerox Foundation on this work.
Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] http://mturk-tracker.com/

minimally acceptable quality and moving to work on the next task than spending more time on the task to produce higher-quality work. Harris (2011) studied performance-contingent financial incentives (both rewards and penalties) and showed that the quality of work was higher when having such incentives than when not having them. Amir, Rand, and Gal (2012) replicated results of some classical lab experiments in MTurk, suggesting that workers in MTurk reacted to performance-contingent financial incentives in a similar way as people in labs. However, neither Harris (2011) nor Amir, Rand, and Gal (2012) cast insights on whether varying the magnitude of performance-contingent financial incentives affects work quality in online labor markets.

Our first goal is to answer the following question:

> *Does the magnitude of performance-contingent financial rewards alone affect work quality and/or worker effort in online labor markets?*

There are reasons to conjecture the answer either way. The intuitive logic for a possibly positive answer is that higher effort, which is costly, will result in improved performance; hence, a worker when offered a higher performance-contingent financial reward is willing to work harder to get a higher reward. On the other hand, the *fair wage-effort hypothesis* (Akerlof and Yellen 1988) in labor economics, which corresponds to the theory of equity (Adams 1963) in social psychology, states that workers have a conception of a fair wage and supply a fraction of their normal effort if the actual wage is less than the fair wage. However, a worker may not have a fair wage for a task a prior and her conception of it may be influenced by a prominent psychological bias, the *anchoring effect* (Tversky and Kahneman 1974; Chapman and Johnson 1994; Ariely, Loewenstein, and Prelec 2003), which refers to the common human tendency to rely heavily on the first piece of information, which may be irrelevant, in making subsequent judgements. In fact, Mason and Watts (2009) found that the perceived appropriate compensation reported by workers in a post-task survey was systematically higher than the payment of the task and monotonically increased with the latter, suggesting that workers used the latter as an anchor. This stream of thinking would suggest a negative answer to the above question. In addition, if the existence of the anchoring effect would make the work quality and worker effort less sensitive to the magnitude of performance-contingent financial rewards, it is natural to ask whether we can improve the effectiveness of such rewards. This leads to our second research question:

> *Can we leverage the anchoring effect in a workflow to improve the effectiveness of performance-contingent financial rewards in online labor markets?*

Our Approach and Results. We design and conduct two experiments in MTurk, one with a task that primarily requires motor skills and the other with a task that demands more cognitive skills. In each experiment, we place two tasks of the same type in each HIT (Human Intelligence Task in MTurk) and consider four levels of performance-contingent financial rewards. These allow us to create three sets of treatments for each experiment: (1) HITs where two tasks have the same reward, (2) HITs where the second task has a higher reward than the first task, and (3) HITs where the second task has a lower reward than the first task.

Our results in the first set of treatments give a negative answer to our first research question: neither work quality nor worker effort is affected by the magnitude of the reward. When comparing the results of the second and third sets of treatments with those of the first set, we find that increasing the reward for the second task leads to higher effort and quality while decreasing the reward for the second task results in lower effort and quality, and the effect is more significant for the motor skill tasks. These results give a positive answer to our second research question. They suggest that the anchoring effect is likely to be an important factor influencing worker behavior in online labor markets and by creating an anchor with an initial reward level, workers may become more sensitive to the magnitude of performance-based financial rewards in subsequent tasks.

Other Related Work. In the context of online labor markets, in addition to work mentioned above, Shaw, Horton, and Chen (2011) compared 14 financial, social, or hybrid incentive schemes, including performance-contingent reward and penalty, in their MTurk experiments. They found that two schemes where a worker's payment depends on the responses of her peers produced higher-quality work. In this paper, we only consider financial rewards that depend on some objective measure of worker performance.

There is a large literature in economics and social psychology on the relationship between financial compensation and productivity, prior to the emergence of online labor markets. Experimental results seem to diverge on this problem. While there is a lot of evidence supporting that higher level of performance-contingent financial incentives improves productivity (Pritchard and Curts 1973; Lazear 2000), some experiments concluded that such incentives had no effect on or even hurt productivity (Jenkins Jr et al. 1998). A well-accepted explanation for financial incentives to hurt productivity is that introducing small financial incentives can decrease intrinsic motivations of workers (Gneezy and Rustichini 2000; Deci, Koestner, and Ryan 1999; Frey and Jege 2001; Bowles 2008) — an effect called *crowding out*. Ariely et al. (2009) also found that overly large performance-contigent rewards hurt performance in a few experiments that required mostly intuitions and simple skills, likely because they triggered overreaction of workers. We refer interested readers to two comprehensive meta-analyses by Camerer and Hogarth (1999) and Jenkins Jr et al. (1998) for more information on this literature.

Many psychological biases have implications on the effectiveness of financial incentives. For example, due to the *loss aversion* effect (Kahneman and Tversky 1984), which refers to people's tendency to strongly prefer avoiding losses to acquiring gains, workers were shown to be more sensitive to wage decrease (Fehr, Goette, and Zehnder 2009). Our work only touches the well-studied anchoring effect. Our experimental design (sequencing two tasks with different reward levels) is inspired by the study of Ariely, Loewenstein, and Prelec (2003), where they showed that in the presence of multiple anchors, the first anchor is most influential.

2 Experimental Design

In our experiments, we place two tasks of the *same type* in each HIT. A worker of a HIT is paid only when she completes both tasks in the HIT. Specifically, each task in the HIT has a performance-independent base payment of 1 cent (i.e. the base payment for the HIT is 2 cents). In addition, each task offers a performance-contingent bonus. This gives us the flexibility of varying the bonus level for tasks in the same HIT. We consider four levels of performance-contingent bonus for individual tasks: 4 cents, 8 cents, 16 cents and 32 cents.

Treatments. We consider the following 10 treatments defined by the bonus level for tasks in the same HIT:

- *4 base treatments*: 4 – 4, 8 – 8, 16 – 16, and 32 – 32;
- *3 treatments with increasing bonus level*: 4 – 8, 4 – 16, and 4 – 32;
- *3 treatments with decreasing bonus level*: 8 – 4, 16 – 4, and 32 – 4.

The 4 base treatments allow us to investigate the effect of the magnitude of performance-contingent rewards on the performance and efforts of workers. The treatments with varying bonuses allow us to create an initial anchor using the bonus of the first task and, when compared with the base treatments, study how the anchoring effect may influence the effect of performance-contingent rewards.

Tasks. To understand whether the effect of performance-contingent rewards depends on the nature of the task, we consider two types of tasks in our experiments.

- *The button clicking (BC) task*: A worker sees a screen with two equal-sized buttons, one on the top and one at the bottom. The "target" button is green, while the other button is gray. The "target" button will alternate between the top button and the bottom button, and the worker is asked to click on the "target" button for as many times as she can in a three-minute task session. The worker receives the pre-specified bonus if she correctly clicks the "target" button for more than 400 times in the session.
- *The spotting differences (SD) task*: A worker is presented two pictures that are identical except at five non-obvious places. The worker is told the number of differences the two pictures have and is asked to find where they differ. Whenever a difference is spotted, the worker can mark it with a red circle by clicking the place in either picture. The worker receives the pre-specified bonus if she correctly spots all five differences. The two spotting differences tasks in a HIT use two different sets of pictures.

The BC task primarily requires motor skills of a worker. A similar task, with buttons placed left and right, was used by Horton and Chilton (2010) in estimating workers' reservation wage in MTurk. The SD task demands more cognitive skills as comparing two pictures engages a worker's short-term memory. Interfaces of both tasks are shown in Figure 1.

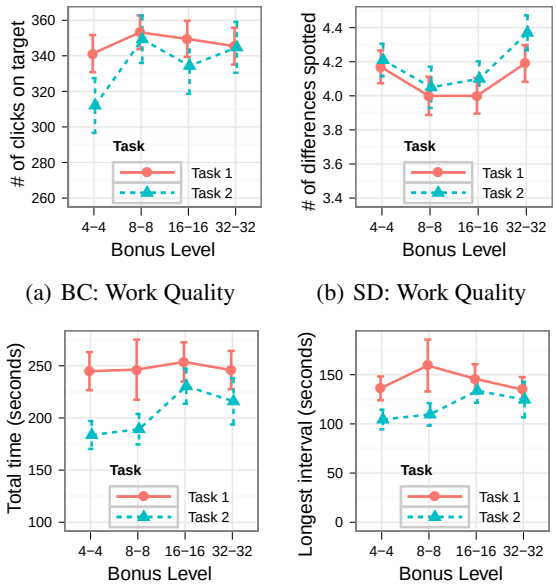
Experimental Control. To avoid complications of culture differences in perceiving financial incentives, we restrict our experiments to U.S. workers. For each type of tasks, each worker is limited to participating in one treatment and working on one HIT so that she is not influenced by other bonus schemes. Upon arrival, a worker is randomly assigned to one of the treatments. For the SD experiment, the two tasks in a HIT are randomly sequenced. We require every worker pass a qualification test on the payment rules after reading the instruction and before proceeding to the actual tasks. Moreover, to ensure that workers pay attention to the possible bonus change, the bonus rule for each task in a HIT is explained immediately before the task, not all together at the beginning of the HIT. After the second task in every HIT, there is a survey asking whether the second task has higher, the same, or lower bonus compared with the first task.

(a) SD task (b) BC task

Figure 1: Interfaces of the Experiments

3 Data

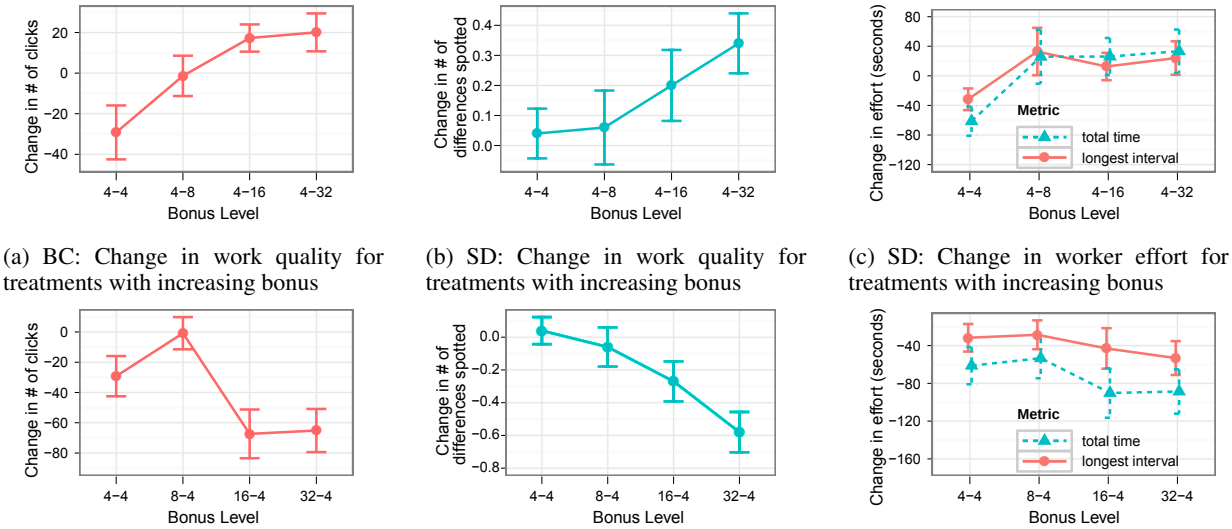
The experiments were conducted over a period of three months. Across all treatments, 1214 workers participated the BC experiment and 1270 workers were recruited on MTurk for the SD experiment. We eliminate all data of those workers who incorrectly answered the bonus comparison question in the survey. For each experiment, we are left with 100 data points for each of the 10 treatments, which we then use in subsequent analyses.



(a) BC: Work Quality (b) SD: Work Quality

(c) SD: Total Time (d) SD: Longest Interval

Figure 2: Work quality and worker effort in base treatments.



(a) BC: Change in work quality for treatments with increasing bonus

(b) SD: Change in work quality for treatments with increasing bonus

(c) SD: Change in worker effort for treatments with increasing bonus

(d) BC: Change in work quality for treatments with decreasing bonus

(e) SD: Change in work quality for treatments with decreasing bonus

(f) SD: Change in worker effort for treatments with decreasing bonus

Figure 3: Changes in Work Quality and Work Effort for Treatments with Changing Bonus Level. Mean values and standard errors of the changes are plotted.

For the BC experiment, we use the number of times a worker clicked the "target" button in the three-minute task session to represent the quality of work produced. For the SD experiment, the number of differences that a worker correctly identified is used to measure the work quality.

While there is no straightforward way to differentiate between the quality of work produced by a worker and her exerted effort in a BC task, we consider two natural metrics for measuring a worker's effort in the SD tasks. We record a worker's activities in a SD task as a sequence of timestamps. For a worker who identified $n \leq 5$ differences correctly in a task, she had a record of $(t_0, t_1, ..., t_n, t_{n+1})$, with t_0 being the time at which she loaded the task page, t_i being the time at which the i-th difference was correctly identified for all $1 \leq i \leq n$, and t_{n+1} representing the time at which she submitted her answer. The first metric for assessing the worker's effort is the *total time* she spent on the task, i.e. $t_{n+1} - t_0$. It captures the duration of her effort. The second metric is the longest elapsed time between two subsequent timestamps, that is, $\max\{t_1 - t_0, \ldots, t_{n+1} - t_n\}$. We call this metric *longest interval*. It captures how hard the worker tried.

4 Results

We first analyze our base treatments to answer the first research question, and then compare work quality and worker effort in treatments with changing bonus levels with those in the base treatments to study our second research question.

Magnitude of Financial Rewards

Figures 2(a) and 2(b) present the mean values of work quality measures in our base treatments for the BC and SD experiments respectively. For the SD experiment, two measures of worker effort (total time and longest interval as defined in Section 3) are presented in Figures 2(c) and 2(d). For both the first and second tasks, both work quality and worker effort have similar mean values across different bonus levels. The first and second tasks exhibit different levels of work quality and worker effort, indicating the possibility of worker learning (or getting bored).

We then test whether the empirical distributions of the work quality and worker effort measures are statistically the same across 4 bonus levels in our base treatments. For example, for the first task in the BC experiment, we have 4 samples of the number of clicks on the target button, one for each of the base treatments and containing 100 data points; we intend to test whether these 4 samples come from the same distribution. The statistical test used is the Kruskal-Wallis one-way analysis of variance (Kruskal-Wallis one-way ANOVA), which is a non-parametric method for testing whether multiple samples originate from the same distribution. The p-values of the test for the first and second tasks are reported in Table 1. The test results indicate that the magnitude of performance-based financial rewards alone does not affect work quality and worker effort.

Table 1: p-values of the Kruskal-Wallis One-Way ANOVA on base treatments.

Metrics	Task 1	Task 2
BC: # of clicks on target	0.82	0.40
SD: # of differences spotted	0.33	0.15
SD: total time	0.37	0.29
SD: longest interval	0.41	0.24

Influencing the Effectiveness of Financial Rewards

Proceeding to examine treatments with changing bonus levels, we focus on exploring whether the bonus for the first task in a HIT may serve as an initial anchor for workers and make them sensitive to the bonus change in the second task.

Table 2: Differences of the mean values for the change in worker quality and worker effort for pairs of treatments. For a pairwise comparison, treatment A vs. treatment B, the reported value for a metric is the mean change of the metric in treatment B minus the mean change in the metric in treatment A. X is fixed to be 4. The statistical significance of the two-sided t-test is marked as a superscript, with †, *, **, and *** representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

(a) Base treatments vs. treatments with increasing bonus

Metrics	X – X vs. X – Y			Y – Y vs. X – Y		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	27.86*	46.50***	49.30***	2.54	32.48*	20.77†
SD: change in # of differences spotted	0.02	0.16	0.30*	0.01	0.10	0.16
SD: change in total time	87.09*	87.35**	94.52**	82.97*	49.38†	63.27†
SD: change in longest interval	64.58*	44.29*	55.76*	82.56*	24.75	34.39*

(b) Base treatments vs. treatments with decreasing bonus

Metrics	Y – Y vs. Y – X			X – X vs. Y – X		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	3.06	-52.15**	-64.46***	28.38	-38.13*	-35.93*
SD: change in # of differences spotted	-0.11	-0.37**	-0.76***	-0.10	-0.31*	-0.62***
SD: change in total time	3.79	-68.66*	-58.62*	7.91	-30.69	-27.37
SD: change in longest interval	21.15	-30.86	-42.87†	3.17	-11.31	-21.50

In this section, unless otherwise specified, our analysis is on the *change* in work quality and worker effort from the first task to the second task in a HIT. That is, for a metric of work quality or worker effort, the change in it for a HIT equals the value of the metric for the second task minus that for the first task in the HIT. Visually, samples of these changes do not deviate from normal distribution. We hence use one-way analysis of variance (one-way ANOVA) and two-sided t-tests, both assuming normal distribution of errors, in the subsequent statistical analysis.

As a base line, we test whether the change in work quality and worker effort are statistically the same across our 4 base treatments. For each of the metrics (change in the number of clicks on target in the BC experiment, change in the number of differences correctly spotted in the SD experiment, change in the total time spent in the SD experiment, and change in the longest interval in the SD experiment), we test whether the samples for the 4 base treatments (4 – 4, 8 – 8, 16 – 16, and 32 – 32) originate from distributions with the same mean. One-way ANOVA, with p-values 0.39, 0.69, 0.63, and 0.51 for the 4 metrics respectively, couldn't reject that they have the same mean, indicating that the change in any of the metrics from the first task to the second task in a HIT is not affected by the magnitude of the rewards alone, which is consistent with the results shown in the previous subsection.

To see how changes in work quality and worker effort in HITs with changing bonus levels compare with those in the base treatments, we plot the changes across different treatments in Figure 3. Figures 3(a), 3(b), and 3(c) present the changes in work quality and worker effort for the 4 – 4, 4 – 8, 4 – 16, and 4 – 32 treatments, i.e. treatments with the same 4 cents bonus for the first task but increasing bonus for the second task. We see a clear upward trend for changes in both work quality and worker effort as the bonus level of the second task increases, except a slight dip for change in longest interval in the 4 – 16 treatment of the SD experiment. Similarly, Figures 3(d), 3(e), and 3(f) plot the changes in work quality and worker effort for the 4 – 4, 8 – 4, 16 – 4, and 32 – 4 treatments, i.e. treatments with the same 4 cents bonus for the second task but the bonus for the first task ranging from 4 cents to 32 cents. The figures show a downward trend for all metrics, except the change in the number of clicks on target in the 8 – 4 treatment of the BC experiment. These suggest that when workers are given the same bonus for the first task, the higher the magnitude of the bonus for the second task, the higher the work quality and worker effort for the second task are; on the contrary, if workers are offered a lower bonus for the second task than for the first task, the larger the decrease in bonus, the lower the work quality and worker effort for the second task are.

We then take a closer look and examine whether these observed differences are statistically significant. We conduct pair-wise comparisons of the treatments and use the two-sided t-test to examine whether the observed changes in work quality and worker effort for the two treatments in the comparison originate from distributions of the same mean. Let X and Y be two bonus levels, X = 4, and X < Y. Table 2(a) presents the pairwise comparisons of treatments X – X and X – Y and those of treatments Y – Y and X – Y, with Y varies from 8, to 16, to 32. The values reported in the table are the differences of the mean values in the corresponding metric for the two treatments in the pair. For example, for the X – X and X – Y comparison for the change in number of clicks on target in the BC experiment, the value reported is the mean value of the change in number of clicks on target in treatment X – Y minus the mean value of the change in number of clicks on target in treatment X – X. The statistical signifi-

cance of a t-test is noted as a superscript. Table 2(b) presents the same data, but for pairwise comparisons of treatments Y – Y and Y – X and those of treatments X – X and Y – X, with Y varies from 8, to 16, to 32.

Thus, each of the pairwise comparisons in Table 2(a) compares a base treatment with a treatment with increasing bonus. If increasing bonus for the second task improves work quality and worker effort on the task, we expect to see positive numbers in this table, which is indeed the case. For the BC task, the improvement in work quality appears to be statistically significant. For the SD task, the improvement in work quality is not statistically significant except when the bonus increase is very large (from 4 to 32), but the increase in worker effort is statistically significant for most treatments with increasing bonus. Similarly, each of the pairwise comparisons in Table 2(b) compares a base treatment with a treatment with decreasing bonus. If decreasing bonus for the second task is detrimental to work quality and worker effort on the task, we expect to see negative numbers in this table, which is mostly true with some exceptions. None of the positive differences are statistically significant. The decrease in work quality for both the BC and SD tasks are statistically significant for larger bonus decreases. We see statistical significance on the decrease in worker effort for some cases of the SD task, but not for the majority of cases. Overall, workers of the BC task response to bonus change in a more sensitive manner than workers for the SD task. A possible reason is that the SD task is more interesting and hence workers have a higher intrinsic motivation to contribute.

Finally, we fit the data to a linear model to show that the absolute magnitude of the reward for the second task does not affect work quality and worker effort on the task, but the change of the magnitude of the reward from the first task to the second task does. The model we use is:

$$M_{i,2} = C + \alpha \cdot M_{i,1} + \beta \cdot \text{Bonus}_{i,2} + \gamma \cdot \Delta\text{Bonus}_i + \varepsilon_i, \quad (1)$$

where M is one of the metrics of work quality or worker effort (i.e. number of clicks on the target button for a BC task, number of differences correctly spotted for a SD task, total time for a SD task, or longest interval for a SD task), $M_{i,1}$ and $M_{i,2}$ are worker i's value of this metric on the first and second tasks respectively, $\text{Bonus}_{i,2}$ is the bonus level of the second task in this HIT, and ΔBonus_i is the change of the bonus level from the first task to the second task, i.e. $\Delta\text{Bonus}_i = \text{Bonus}_{i,2} - \text{Bonus}_{i,1}$. Note that here we consider the value of a metric, rather than the change in the value of a metric. We include $M_{i,1}$ in the model to account for a worker's innate capability on the task. The regression results for all four metrics are shown in Table 3. It is clear that the bonus level of the second task does not affect either work quality or worker effort, but the change of the bonus level affects both.

Table 3: Regression results for linear model (1). Estimated coefficients and standard errors are reported. The statistical significance is marked as a superscript, with *, **, and *** representing significance levels of 0.05, 0.01, and 0.001 respectively.

Metric M	C	$M_{i,1}$	$\text{Bonus}_{i,2}$	ΔBonus_i
# of clicks on target in BC	48.2^{**} (14.8)	0.81^{***} (0.04)	0.23 (0.48)	1.66^{***} (0.36)
# of differences spotted in SD	1.79^{***} (0.14)	0.55^{***} (0.03)	0.004 (0.004)	0.01^{***} (0.003)
Total time in SD	168.0^{***} (14.8)	0.18^{***} (0.03)	0.48 (0.82)	1.87^{**} (0.63)
Longest interval in SD	110.3^{***} (11.1)	0.11^{**} (0.03)	0.16 (0.66)	1.15^{*} (0.50)

5 Conclusion

We investigate the effects of performance-contingent financial rewards in online labor markets and attempt to provide answers to two research questions: (1) does the magnitude of performance-contingent financial rewards alone affect work quality and/or worker effort in online labor markets? (2) can we leverage the anchoring effect in a workflow to improve the effectiveness of performance-contingent financial rewards in online labor markets? We give a negative answer to the first question and a positive answer to the second question by conducting two experiments in MTurk, one with a task that primarily requires workers to use their motor skills and the other with a task that demands more cognitive skills.

It is shown that the magnitude of performance-contingent financial rewards alone does not affect work quality and worker effort. However, the change of the bonus level from the first task to the second task in a HIT significantly affects both of them on the second task — increasing the bonus improves them while decreasing the bonus hurts them. Our results support that, given a type of tasks, workers in MTurk may use the payment of the first task of the type that they encounter as an anchor to form their conception of a fair payment for this type of tasks and their behavior is consistent with the conjecture of the fair wage-effort hypothesis.

The practical implication of our results is that the design of crowdsourcing workflows can possibly affect the effectiveness of incentives and hence the performance of workers. Such impacts should be taken into consideration in designing workflows.

The results of our experiments support the observation of Ariely, Loewenstein, and Prelec (2003) on the importance of the initial anchor. As a future direction, we would like to explore the impact of subsequent anchors on the effectiveness of financial rewards. For example, when we have HITs with more than two tasks and varying bonus levels (not necessarily monotonic), do bonus levels for all previous tasks affect work quality and worker effort for the current task? Another direction we would like to pursue is to study the effects of various peer prediction methods (Miller, Resnick, and Zeckhauser 2005; Prelec 2004), when combined with financial incentives, on work quality and worker effort. These methods reward a worker based on not only her answer but also the answers of her peers, and hence are especially suitable for crowdsourcing tasks where the work quality is not verifiable or too costly to be practical to verify.

References

Adams, S. J. 1963. Toward an understanding of inequity. *Journal of Abnormal and Social Psychology* 67:422–436.

Akerlof, G., and Yellen, J. 1988. Fairness and unemployment. *The American Economic Review* 78(2):44–49.

Amir, O.; Rand, D. G.; and Gal, Y. K. 2012. Economic games on the internet: The effect of $1 stakes. *PLoS ONE* 7(2):e31461.

Ariely, D.; Gneezy, U.; Loewenstein, G.; and Mazar, N. 2009. Large stakes and big mistakes. *Review of Economic Studies* 76(2):451–469.

Ariely, D.; Loewenstein, G.; and Prelec, D. 2003. Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics* 118(1):73–106.

Benkler, Y. 2002. Coase's Penguin, or, Linux and the Nature of the Firm. *Yale Law Journal* 112(3):367–445.

Bowles, S. 2008. Policies Designed for Self-Interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments Science. 320(5883):1605–1609.

Camerer, C. F., and Hogarth, R. M. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19(1-3):7–42.

Chapman, G. B., and Johnson, E. J. 1994. The limits of anchoring. *Journal of Behavioral Decision Making* 7(4):223–242.

Dai, P.; Mausam; and Weld, D. S. 2010. Decision-theoretic control of crowd-sourced workflows. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, AAAI '10, 1168–1174.

Dai, P.; Mausam; and Weld, D. S. 2011. Artificial intelligence for artificial artificial intelligence. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI '11, 1153–1159.

Deci, E.; Koestner, R.; and Ryan, R. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125(6):692–700.

Fehr, E.; Goette, L.; and Zehnder, C. 2009. A behavioral account of the labor market: The role of fairness concerns. *Annu. Rev. Econ.* 1(1):355–384.

Frey, B. S., and Jege, R. 2001. Motivation crowding theory: A survey of empirical evidence. *Journal of Economic Surveys* 15(5):589–611.

Gneezy, U., and Rustichini, A. 2000. Pay enough or don't pay at all. *The Quarterly Journal of Economics* 115(3):791–810.

Harris, C. 2011. You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 15–18.

Horton, J. J., and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, EC '10, 209–218. New York, NY, USA: ACM.

Jenkins Jr, G.; Mitra, A.; Gupta, N.; and Shaw, J. 1998. Are financial incentives related to performance? a meta-analytic review of empirical research. *Journal of Applied Psychology* 83(5):777.

Kahneman, D., and Tversky, A. 1984. Choices, values and frames. *American Psychologist* 39(4):341–350.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, 467–474. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Lazear, E. P. 2000. Performance pay and productivity. *The American Economic Review* 90(5):1346–1361.

Lin, C.; Mausam; and Weld, D. 2012a. Crowdsourcing control: Moving beyond multiple choice. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, UAI '12, 491–500.

Lin, C.; Mausam; and Weld, D. 2012b. Dynamically switching between synergistic workflows for crowdsourcing. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI '12.

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, 77–85. New York, NY, USA: ACM.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The Peer-Prediction Method. *Management Science* 51:1359–1373.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306:462–466.

Pritchard, R. D., and Curts, M. I. 1973. The influence of goal setting and financial incentives on task performance. *Organizational Behavior and Human Performance* 10(2):175 – 183.

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 *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

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, 275–284. New York, NY, USA: ACM.

Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.