Incentivizing High Quality Crowdwork

CHIEN-JU HO
Cornell University
and
ALEKSANDRS SLIVKINS
and
SIDDHARTH SURI
and
JENNIFER WORTMAN VAUGHAN
Microsoft Research

We study the causal effects of financial incentives on the quality of crowdwork. We focus on *performance-based payments* (PBPs), bonus payments awarded to workers for producing high quality work. We design and run randomized behavioral experiments on the popular crowdsourcing platform Amazon Mechanical Turk with the goal of understanding *when*, *where*, and *why* PBPs help, identifying properties of the payment, payment structure, and the task itself that make them most effective. We provide examples of tasks for which PBPs do improve quality. For such tasks, the effectiveness of PBPs is not too sensitive to the threshold for quality required to receive the bonus, while the magnitude of the bonus must be large enough to make the reward salient. We also present examples of tasks for which PBPs do not improve quality. Our results suggest that for PBPs to improve quality, the task must be *effort-responsive*: the task must allow workers to produce higher quality work by exerting more effort. We also give a simple method to determine if a task is effort-responsive *a priori*. Furthermore, our experiments suggest that all payments on Mechanical Turk are, to some degree, *implicitly* performance-based in that workers believe their work may be rejected if their performance is sufficiently poor. In the full version of this paper, we propose a new model of worker behavior that extends the standard principal-agent model from economics to include a worker's subjective beliefs about his likelihood of being paid, and show that the predictions of this model are in line with our experimental findings. This model may be useful as a foundation for theoretical studies of incentives in crowdsourcing markets.

Categories and Subject Descriptors: H.3.5 [**Online Information Services**]: Web-Based Services; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Economics, Experimentation

Additional Key Words and Phrases: Crowdsourcing; Performance-Based Payments; Incentives

1. INTRODUCTION

Crowdsourcing markets are platforms on which workers around the world perform tasks for pay. In a crowdsourcing market like Amazon Mechanical Turk, requesters post tasks along with the payment amount. Workers can then browse the available tasks and choose tasks to work on.

The full version of this paper appeared in the 24th International World Wide Web Conference (WWW 2015) and is available on `arXiv:1503.05897`
Authors' addresses: `ch624@cornell.edu`, `slivkins@microsoft.com`, `suri@microsoft.com`, `jenn@microsoft.com`

Crowdsourcing markets are used to conduct user studies [Kittur et al. 2008], run behavioral experiments [Horton et al. 2011; Mason and Suri 2012], collect data [Horton and Chilton 2010; Wah et al. 2011], test or even build business applications [Schall 2012; Alonso 2013], and more. While these markets are effective at recruiting diverse labor pools, the quality of work produced varies widely across tasks and workers. The prevalence of low quality crowdwork has inspired a growing literature on techniques to boost accuracy, for example, by using redundant assignments for labeling tasks [Sheng et al. 2008; Ipeirotis et al. 2010; Karger et al. 2011; Liu et al. 2012; Ho et al. 2013], smartly assigning tasks to workers [Ho and Vaughan 2012; Ho et al. 2013], introducing social incentives [Rogstadius et al. 2011; Shaw et al. 2011], or altering financial incentives [Mason and Watts 2009; Rogstadius et al. 2011; Buhrmester et al. 2011; Shaw et al. 2011; Harris 2011; Yin et al. 2013; 2014; Gilchrist et al. 2014]. These solutions have had mixed success, and how to improve the quality of work in general is still not well understood.

In this paper, we study the use of financial incentives to encourage high quality crowdwork on Amazon Mechanical Turk. In particular, we focus on the use of *performance-based payments* (PBPs), bonus payments awarded to workers for producing high quality work. Previous empirical studies of performance-based payments in crowdsourcing markets have produced mixed and somewhat contradictory recommendations. Harris [2011] and Yin et al. [2014] suggested that PBPs can improve work quality, while Shaw et al. [2011] found no improvement and Yin et al. [2013] found no difference in quality when varying bonus size.

Our results explain these disparities in prior work. Furthermore, we show how to generalize previous findings beyond the particular tasks that were studied. We design and run experiments with the goal of understanding not just whether PBPs improve work quality for a specific task or bonus size, but *when*, *why*, and *where* they improve work quality. We identify properties of the payment, payment structure, and the task itself that make PBPs effective.

2. DOES PBP WORK?

Workers were asked to proofread an article and correct spelling errors. For each article, we randomly inserted 20 typos from a list of common spelling errors. Workers were asked to input the line number of each typo, the misspelled word, and the correct spelling of the word.

This task has two key properties. First, we would expect that workers could produce better work by exerting more effort—the more carefully a worker reads or the more passes a worker takes over the text, the more typos he will find—and that this would open up the possibility of PBPs improving quality. (We study this conjecture in more detail in Section 4.) Second, since we injected the typos into the text, the quality of each worker's output could be measured objectively, though this was not known to the workers.

After workers accepted the task (HIT), they were randomly assigned to different treatments and then shown treatment-specific instructions, when applicable. Our experiment had a 2×3 design, with 2 treatments governing the base payment and 3 treatments governing the bonus payment (if any). We discuss the bonus treatments first:

- *No Bonus:* This is the control group. It had no bonus and no mention of a bonus.
- *Bonus for All:* All workers earned a $1 bonus after submitting the HIT.
- *PBP:* Workers earned a $1 bonus if they found 75% of the typos found by the other workers.

We are also interested in whether workers have subjective assumptions on how much effort they must exert to get their work accepted. Workers may be afraid that if they do not find a sufficient number of typos their work will be rejected, resulting in no pay and a negatively affected MTurk reputation. To estimate this, we designed a treatment in which workers were explicitly guaranteed acceptance provided that they completed a very small amount of work. We had two treatments for the base payment:

- *Non-Guaranteed:* There were no extra instructions. This is the control and emulates most MTurk tasks.
- *Guaranteed:* Workers were told they would get paid if they found at least one typo.

The first typo appeared before line 3 in each article. Thus a worker would only have to do a trivial amount of work to ensure they got paid in the guaranteed base treatment.

2.1 Results

The HIT was completed by 1,000 unique workers, who were each assigned uniformly to one of the six treatments. The primary dependent variable was the number of true typos found. In the analysis we made six comparisons that we spell out below. We performed this analysis using an analysis of variance (ANOVA) with one-sided, planned comparisons [Seltman 2014] and report p-values that have been corrected for these multiple (six) comparisons. The results of this experiment are shown in Figure 1 and described below.

PBPs improve quality. To determine whether PBPs increase quality for this task, we focus on the non-guaranteed base treatments since almost all HITs on MTurk do not explicitly guarantee any kind of acceptance criteria. Workers in the PBP bonus treatment found on average 1.3 more typos than workers in the No Bonus treatment ($p = 0.042$), showing that PBPs did improve quality for this task.

All payment schemes may be implicitly performance-based. In the No Bonus treatment, the guaranteed base resulted in 1.5 fewer typos found on average compared with the non-guaranteed base ($p = 0.015$). Similarly, in the Bonus for All treatment, the guaranteed base resulted in 1.3 fewer typos found on average ($p = 0.024$). While there may be other explanations, this suggests that workers do have subjective beliefs on the amount of work that needs to be done for their work to be accepted, lending support to our conjecture that payments are already implicitly performance-based.

In the PBP bonus treatment, we did not see a significantly different effect between the guaranteed base and non-guaranteed base treatments. We offer two related explanations of this finding. First, the only way to grant a bonus using the MTurk

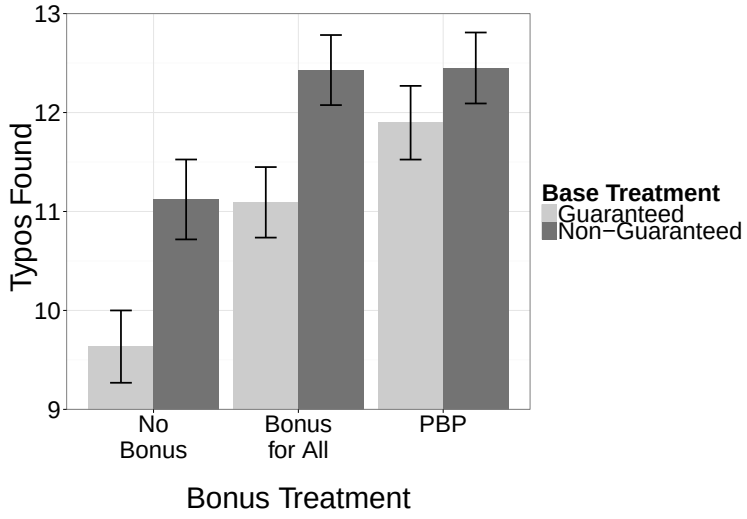


Fig. 1. The effect of different payment schemes on work quality in the proofreading task. Error bars indicate the mean ± one standard error.

API is to first accept the work. This means that in the PBP bonus treatment, workers would likely believe that finding 75% of typos would almost certainly result in their work being accepted, already altering their subjective beliefs. Second, the treatment might have made this 75% threshold more salient to the workers. This gave a clear goal for the workers to strive for.

Simply paying more improves quality. Focusing again on the non-guaranteed base treatment, workers in the Bonus for All treatment found on average 1.3 more typos than workers in the No Bonus treatment ($p = 0.036$). Thus offering an unconditional bonus—which is essentially just paying more—increased quality.

This finding is perhaps surprising since it appears to contradict the results of prior work [Mason and Watts 2009; Rogstadius et al. 2011; Buhrmester et al. 2011]. We give two potential explanations. First, since the announcement of the bonus came after workers accepted the HIT, the workers may be exhibiting reciprocity by doing higher quality work [Gilchrist et al. 2014], rewarding the requester for this pleasant surprise. We further test and refute this hypothesis in Section 3. Second, this could be explained by the implicit PBP effect described above. That is, workers might have subjective beliefs about the number of typos they must find to get paid. If we increase the bonus payment, workers might be willing to put in more effort to increase their probability of earning this higher amount.

This observation is not inconsistent with previous work. In most prior work, either easy tasks were chosen which might cause workers to perform well even for low pay [Rogstadius et al. 2011; Buhrmester et al. 2011] or additional instructions or tutorials were provided which may have primed workers' subjective beliefs [Mason and Watts 2009].

PBPs can save money compared with high unconditional payments. In the non-guaranteed base treatment, the difference in the number of typos found in

the PBP and Bonus for All treatments is not significant. Both resulted in higher quality work than the control. However, we spent much less money on the PBP treatment. We paid each worker $1.50 in the Bonus for All treatment, while we paid each worker only $0.97 on average in the PBP treatment with non-guaranteed base and $0.96 on average in the PBP treatment with guaranteed base. Therefore, it may still be advantageous for requesters to offer PBPs even if they could achieve the same quality work with unconditional payments.

Having established that PBPs can improve quality for the proofreading task, we investigated the effect of varying two parameters of the payment scheme: the bonus threshold and the bonus amount, to better understand when PBPs help. Due to space restrictions we omit the description of this round of follow-up experiments. Our results indicate that PBPs improve quality for a wide range of possible thresholds, provided that the requester offers a bonus that is high enough to make the extra reward salient. More specifically, if the bonus offered is too small, PBPs do not improve quality (and can even reduce quality), which could explain why Shaw et al. [2011] reported little or no quality improvement using PBPs compared with fixed payments. Additionally, we found diminishing returns from increases in the payment beyond a certain point, which could explain why Yin et al. [2013] found that bonus size had little effect on quality.

3. WHY DOES PBP WORK?

There are two primary motivations for our next experiment. First, we wanted to verify that PBPs are useful in other tasks beyond finding typos. Second, we wanted to explore potential reasons why PBPs work. In particular, as pointed out in Section 2.1, simply increasing the amount of the bonus payment led to almost as much of an improvement as using PBPs in the proofreading experiment. While it could be that workers are responding rationally to the provided incentives, it could also be the case that workers are increasing their effort due to a reciprocity effect; workers are pleasantly surprised to discover the opportunity to receive a (performance-based or unconditional) bonus after accepting the HIT, and reward the requester for this kind action by working harder. Indeed, Gilchrist et al. [2014] found, in a different crowdsourcing context, that workers who accept a task and then receive an unexpected bonus do higher quality work than workers who are paid the same amount total but are told up front. This experiment is designed to test whether this "unexpected bonus effect," is the (partial) cause of the observed increases in performance using PBPs.

3.1 Experiment Design

In this task, workers were shown twenty pairs of images. Ten of the pairs were identical images, while the other ten pairs contained minor differences. Workers were asked to specify whether each pair was identical or not, and were not told how many pairs of images were identical in advance. Again, this task has two key properties we desire. First, we speculated that workers would be more likely to spot the differences between images if they spent more time and effort looking. Second, we can objectively measure the quality of workers' output by the number of correctly answered pairs. A similar task was used in experiments by Yin et al. [2013]. We next describe the treatments:

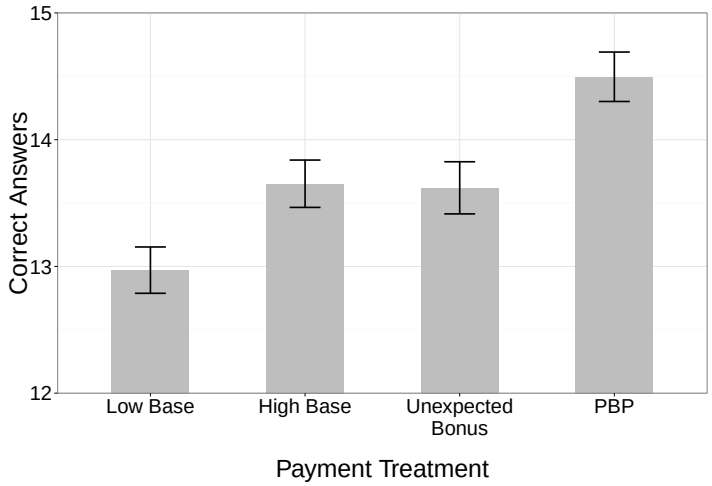


Fig. 2. The effect of different payment schemes on work quality in the spot the differences task. Error bars indicate the mean ± one standard error.

- *Low Base:* The base payment was \$0.50. No opportunity for a bonus was given. This was our control.
- *High Base:* The base payment was \$1.50. No opportunity for a bonus was given.
- *Unexpected Bonus:* The base payment was \$0.50. After accepting the HIT, workers were told they would receive an additional bonus of \$1.
- *PBP:* The base payment was \$0.50. In addition to the base payment, workers could earn a bonus of \$1 if they correctly labeled 80% of the image pairs as identical or not. Workers were informed of the bonus and rules for receiving the bonus before accepting the HIT.

Note that the payment amounts in the High Base and Unexpected Bonus treatments are the same. The difference is only how and when the payments were described.

3.2 Results

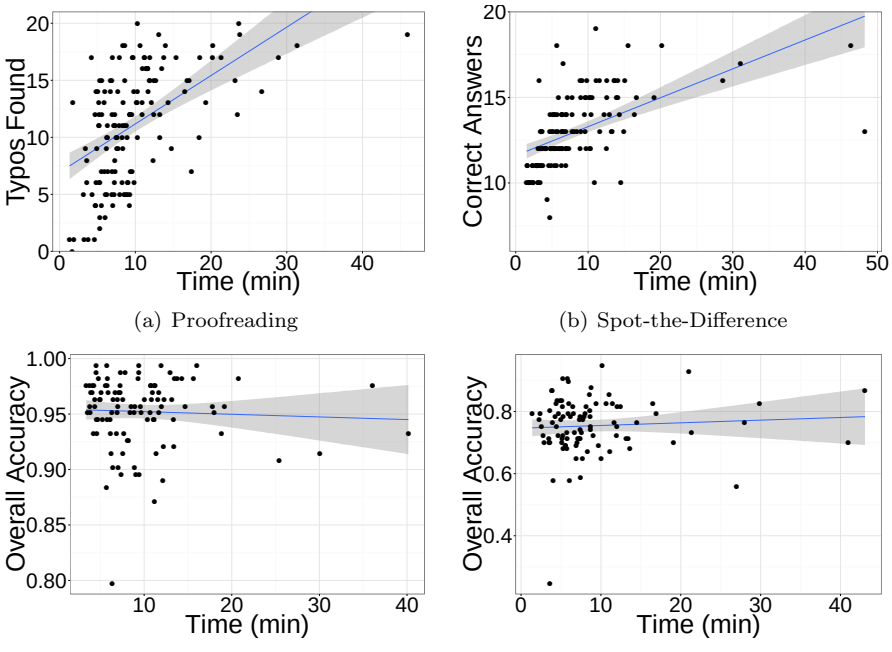
To avoid selection bias in workers accepting HITs with varying pay rates we randomly chose 800 workers from a pool that completed a qualification HIT and randomly assigned them to the four treatments, 200 workers each. After assigning qualifications corresponding to each treatment, we posted the HITs for each simultaneously and sent each worker a notification with a link to their treatment's HIT. We conducted a chi-squared test to check for significant differences in the number of participants finishing the four treatments and found none ($p = 0.90$). In the analysis (see Figure 2), we make six comparisons, described below. We did this analysis using an ANOVA with one-sided, planned comparisons [Seltman 2014] and report p-values that have been corrected for these multiple comparisons.

Similar to the proofreading experiment described in Section 2, simply paying more resulted in higher quality work. The High Base treatment had a significantly

higher number of correct answers than the Low Base treatment ($p = 0.030$). Similarly, the Unexpected Bonus treatment had a significantly higher number of correct answers than the Low Base treatment ($p = 0.047$). Figure 2 shows no significant difference between the High Base and the Unexpected Bonus treatments. This suggests that there was no "unexpected bonus effect" in contrast to Gilchrist et al. [2014].The absence of any reciprocity effect due to the unexpected bonus suggests that workers were doing better work to increase the probability (according to their prior assumptions) that their work got accepted and thus earn the higher pay.

We also observe that workers in the PBP treatment outperformed workers in all other treatments ($p < 0.005$). This suggests that workers are rational to some degree and are willing to exert more effort to increase their chances of receiving higher payments. Note that in this experiment workers knew *before* they accepted the HIT that they could earn a bonus, in contrast to the experiment described in Section 2 in which workers were informed of the opportunity to earn a bonus only *after* they accepted the HIT. We have therefore shown that PBPs can work whether or not the opportunity for a bonus is expected.

4. WHERE DOES PBP WORK?



(a) Proofreading (b) Spot-the-Difference

(c) Handwriting Rec. (d) Audio Transcription

Fig. 3. Time vs. quality for effort responsive tasks in panels 3(a) and 3(b), and non-effort responsive tasks in panels 3(c) and 3(d). The blue lines indicate the regression line and the shaded areas represent the 95% confidence interval around it. Results are similar when outliers are excluded from the analysis.

We have shown that PBPs incentivize higher quality crowdwork on two specific tasks, proofreading and spotting differences in images. It is natural to ask whether our results generalize, and in particular, what properties of a task open up the possibility of performance improvements with PBPs.

Camerer and Hogarth [1999] note that in the context of economics lab experiments, performance-based incentives tend to improve quality for *effort-responsive* tasks, tasks for which it is possible to generate higher quality work by exerting additional effort (presumably without requiring *too much* effort). One might ask if the same is true in a crowdsourcing setting. Since it is difficult to directly measure how much effort a worker has put into a task, we use the time a worker spent on a HIT as a proxy measure for effort, and examine the relationship between time spent and quality of work.

Figures 3(a) and 3(b) illustrate the correlation between work quality and time for the proofreading and spot-the-difference tasks respectively. Each shows the amount of time that a worker spent on the HIT versus the quality of his work. We see that, in general, workers who spent more time on our tasks generated better quality work. We observe similar trends in all treatments, but include only workers in the control groups in the plots since they are most comparable across tasks. This is evidence that the tasks on which we observed improvements from PBPs are effort-responsive.

To further explore this hypothesis and the generalizability of our results, we examined the effects of PBPs on two additional tasks, handwriting recognition and audio transcription, which is one of the most common tasks on MTurk. We did treatments with PBPs and control treatments (with no bonus) for both tasks. Again, due to space restrictions we omit the details of these follow up experiments. Figures 3(c) and 3(d) show that the quality of work produced was not significantly correlated with the time a worker spent for either task. In other words, neither task appears to be effort-responsive. Moreover, we did not find a significant difference between the accuracy of workers in the control groups versus the PBP treatments in either task via one-sided t-tests.

A Practical Recommendation.

While the results in this section are not causal, they are in line with the hypothesis that the extent to which a task is effort-responsive is an important reason for whether PBPs help improve quality for this task. This suggests an approach that requesters can use when deciding whether to employ PBPs in their own HIT. A requester could run a pilot of their HIT with a small number of workers and a fixed (not performance-based) payment and plot the time that workers spend on the task versus the quality of their work to determine whether and to what extent the task is effort-responsive. A requester may be able to incentivize higher quality using PBPs only if the task is (sufficiently) effort-responsive. In this case, the requester must determine whether the boost in quality is worth the extra cost of PBPs.

REFERENCES

Alonso, O. 2013. Implementing crowdsourcing-based relevance experimentation: An industrial perspective. *Information Retrieval 16,* 2, 101–120.

Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*.

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

Gilchrist, D., Luca, M., and Malhotra, D. 2014. When 3+1 > 4: Gift structure and reciprocity in the field. Tech. rep. Working Paper.

Harris, C. G. 2011. You're hired! An examination of crowdsourcing incentive models in human resource tasks. In *WSDM 2011 Workshop on Crwdsourcing for Search and Data Mining*.

Ho, C.-J., Jabbari, S., and Vaughan, J. W. 2013. Adaptive task assignment for crowdsourced classification. In *ICML*.

Ho, C.-J. and Vaughan, J. W. 2012. Online task assignment in crowdsourcing markets. In *AAAI*.

Horton, J. J. and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *ACM EC*.

Horton, J. J., Rand, D., and Zeckhauser, R. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics 14,* 3, 399–425.

Ipeirotis, P. G., Provost, F., and Wang, J. 2010. Quality management on Amazon Mechanical Turk. In *HCOMP*.

Karger, D., Oh, S., and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems. In *NIPS*.

Kittur, A., Chi, E., and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI*.

Liu, Q., Peng, J., and Ihler, A. 2012. Variational inference for crowdsourcing. In *NIPS*.

Mason, W. and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods 44,* 1, 1–23.

Mason, W. and Watts, D. J. 2009. Financial incentives and the "performance of crowds". In *HCOMP*.

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 *ICWSM*.

Schall, D. 2012. *Service-Oriented Crowdsourcing - Architecture, Protocols and Algorithms*. Springer Briefs in Computer Science. Springer.

Seltman, H. J. 2014. Experimental design and analysis. `http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf`.

Shaw, A. D., Horton, J. J., and Chen, D. L. 2011. Designing incentives for inexpert human raters. In *CSCW*.

Sheng, V., Provost, F., and Ipeirotis, P. 2008. Get another label? Improving data quality using multiple, noisy labelers. In *KDD*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology.

Yin, M., Chen, Y., and Sun, Y.-A. 2013. The effects of performance-contingent financial incentives in online labor markets. In *AAAI*.

Yin, M., Chen, Y., and Sun, Y.-A. 2014. Monetary interventions in crowdsourcing task switching. In *HCOMP*.