You Get What You Pay For: Experimental Analysis on the Relationship Between Pay and Work Quality

Legg Yeung, Stanimir Vichev & Frederic Suares

December 3, 2017

```
# load packages
library(foreign)
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.3.3
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.3.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(data.table)
library(multiwayvcov)
```

```
## Warning: package 'multiwayvcov' was built under R version 3.3.3
```

Introduction

In most economies, it is generally believed that renumeration for someone's work is strongly related to the effort they will put into it, and the eventual quality of the results. Unfortunately, this is a concept that is challenging to test in the normal world. Employers cannot easily conduct experiments with their own employees, say, by giving them similar tasks and different payments on a random basis, as this could be considered unethical and would lead to a serious disruption in the workforce. At the same time, such a study would be very helpful to employers who want to understand what motivates their employees, and what part the pay plays. The Amazon Mechanical Turk (AMT) platform for the crowdsourced completion of tasks provides a great opportunity for experimentally testing the relationship between payment and quality of work without having to worry about subject interaction or high costs of wasted man-hours. Our experiment uses the AMT platform to experimentally test whether payment for a task has a positive effect on the quality of its result. We used two different experimental designs (normal and stepped wedge), randomly putting subjects in different payment brackets and measuring how results differed between groups. The paper is organized as follows: section 2 discusses background and motivations, section 3 explains the experimental design, detailing the platforms used and the experimental schedule, sections 4 and 5 present the data and analysis for the two types of experiments, section 6 discusses overall results, section 7 looks at possible future studies, followed by a conclusion and bibliography.

Related Work

The use of online labour markets as an effective and efficient platform for social science experimentation has been noted by several studies, and explored in detail in Horton et. al. 2011. They perform several successful experiments and even look at the labour supply curves of workers. This shows that we have made the right choice of platform to conduct our research. Another experiment done by Horton & Chilton, 2010, develops a novel method for estimating the smallest price for a task that a worker would accept. They also look into the way workers respond to incentives, with some being rational and some setting earnings targets. Finally, Mason & Watts, 2009, use the AMT platform to explore the effect of financial incentives on the performance of workers. They conclude that higher financial incentives increase the quantity, but not quality, of the work done by workers, citing an anchoring effect as the cause of this. By doing a similar experiment nearly 9 years later, we hope to see whether we get the same results as online labour markets such as AMT gain more prominence and popularity, leading to a more diverse market with more workers and requestors.

Experimental Design

The AMT platform allows us, as requestors to create a task that is of a difficulty of our choosing, which we could then post on the platform, defining how much we are willing to pay and how many people we want doing it. AMT workers (our subjects) see the task and how much it pays and decide whether or not they want to take part. Once the task is completed by all the workers, it gets taken off the platform and we can download and analyze the results. Through AMT's qualification system we could make sure that each worker only took part in a single task. By using their worker IDs and randomly assigned survey IDs, we were able to track which workers actually finished the task and which just looked at it.

The task we set for this experiment consisted of a survey (created and managed using Qualtric) that asked people to complete 8 personal details questions (knowledge of dogs, preference for work with certain media, etc.), followed by 50 multiple choice questions. Each multiple choice question asked people to look at a photo of a dog and pick the breed it belongs to out of 8 possible choices. The workers were also given a document presenting a sample picture for each breed, so that workers' results depended only on their efforts and not on their general knowledge of dogs, since that would have otherwise lead to skewed results. To capture the quality of work, we measured the accuracy with which workers correctly classified images. Two screener questions were included that asked workers to classify pictures of cats, so that we could isolate workers that went through the survey without reading the questions.

For this study we followed two experimental designs: a between-subjects design and a stepped wedge design, both of which used the same task to measure accuracy. The first set of experiments followed a between-subjects design, and involved releasing the task on the AMT platform at different times with one of four possible payments ($0.10, $0.25, $0.40, $0.55). We released only one task at a time (otherwise workers would only do the higher paying one), waiting for 100 workers to complete the task before pulling it off the platform. Through Qualtric, which hosted the survey, we could track people that only looked at the survey without finishing it, allowing us to exclude them from future tries. Apart from tracking how workers answered all the questions, Qualtric allowed us to track how long they spent doing the survey, which we also consider in our analysis. One important thing to note about this approach is that it suffers from the downside that different types of people may choose to take part in tasks that present different payments, so we may not always be comparing apples to apples when looking at different groups. However, we hope that AMT's population of workers is large enough to provide enough random sampling (**this needs to be fixed and explained better**).

(Add table here) - col1: Name of Experiment (e.g. Design 1, Order 1, $??) - col2: $ Treatment - col3: The PST date time this Mturk posting was posted - col4 : How long it took for all 100 HITs to be submitted - col5: How many completed HITs were received in the first 30min after launching. - col6: Average accuracy for that posting

The second set of experiments involved following a stepped wedge design, where we released a single task

on the AMT platform of 48 questions to be done by 400 workers that pays $0.10. Once a worker starts the survey, we would randomly assign her to one of four treatment groups, using functionality provided by Qualtric. The first group would complete the 48 questions without any difference to the first design. The second group would complete the first 32 questions normally, but they would then be told that they would actually be paid $0.15 extra before doing the final 16 questions. The third group would be told about the bonus after completing the first 16 questions, and the fourth group would be told about the bonus before starting any of the questions. This way we hope to achieve real random assignment and truly compare apples to apples.

```
d2df = data.frame(Wave=c("A", "B", "C", "D"), Step_1_Pay=c(0,0,0,0.15), Step_2_Pay=c(0,0,0.15,0.15), St
print(d2df) # Look for better formatting
```

```
##   Wave Step_1_Pay Step_2_Pay Step_3_Pay
## 1    A       0.00       0.00       0.00
## 2    B       0.00       0.00       0.15
## 3    C       0.00       0.15       0.15
## 4    D       0.15       0.15       0.15
```

```
knitr::kable(d2df, "markdown")
```

Wave	Step_1_Pay	Step_2_Pay	Step_3_Pay
A	0.00	0.00	0.00
B	0.00	0.00	0.15
C	0.00	0.15	0.15
D	0.15	0.15	0.15

AMT and Qualtric

Experiment Schedule

Pilot Experiment

Design 1 Experiments

Data

Analysis

Design 2 Experiments

Data

Analysis

Combined Results and Analysis

Future Work

Conclusion

Bibliography

Mason, Winter and Watts, Duncan J. 2010. Financial incentives and the "performance of crowds". SIGKDD Explor. Newsl. 11, 2 (May 2010), 100-108. DOI=http://dx.doi.org/10.1145/1809400.1809422 . https://dl.acm.org/citation.cfm?id=1809422

Horton, J.J., Rand, D.G. & Zeckhauser, R.J. Exp Econ (2011) 14: 399. DOI=https://doi.org/10.1007/s10683-011-9273-9 . https://link.springer.com/article/10.1007%2Fs10683-011-9273-9?LI=true

Horton, J.J., Chilton, L.B.2011. The Labor Economics of Paid Crowdsourcing. https://arxiv.org/pdf/1001.0627.pdf

```
y_contr = c(3,4,5,4,3,2)
y_treat = c(2,3,4,8,6,4)
comply = c(1,2,3)
ate = mean(y_treat - y_contr)
cace = mean(y_treat[comply]-y_contr[comply])
cat("ATE:",ate,"\n","CACE:",cace,"\n")
```

```
## ATE: 1
##  CACE: -1
```