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

Legg Yeung, Stanimir Vichev, Frederic Suares

University of California, Berkeley

December 17, 2017

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean a sem vitae mi lobortis condimentum. Maecenas rutrum vitae libero in tempor. Fusce tempor mauris nec vulputate luctus. Phasellus et semper nisi, sed luctus erat. Duis eu congue lectus, ac mattis diam. Fusce tristique efficitur lacus, vel vehicula leo mollis id. Proin vel venenatis velit. Sed et semper ante. Interdum et malesuada fames ac ante ipsum primis in faucibus. Sed id mi urna. Integer urna nunc, feugiat sed felis vitae, imperdiet tristique ipsum. Sed pretium eros massa, a sollicitudin diam finibus aliquet.

Fix section numbers in the introduction

Add optional table of contents here

1 Introduction

In most economies, it is generally believed that remuneration 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 higher payment for a task has a positive effect on the quality of its result. We used two different experimental designs (traditional between-subject and stepped-wedge), one randomly assign tasks to four different payment levels and the other randomly assign turkers to four different payment levels, to measure how resultant quality of work differ 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 experiment designs, section 6 discusses overall results, section 7 looks at possible future studies, followed by a conclusion and bibliography.

2 Related Work

The use of online labor 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 labor 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 labor markets such as AMT gain more prominence and popularity, leading to a more diverse market with more workers and requestors.

3 Research Hypothesis, Identification Strategy

We hypothesis that higher payment per human intelligence task (HIT) on average would lead to higher task performance. To operationalize this construct, we define the treatment variable as pay rate in US dollars, and outcome variable as proportion of image classification questions scored correctly in each returned HIT. In each of the two experimental designs, four different pay rates are randomly assigned to each HIT. Similarly, in each of the two experimental designs, a total of 50 image classifications questions on dog breeds are prompted in each HIT. The four pay rates are chosen between \$0.10 and \$0.55, which correspond to the lower and upper bound we commonly see for similar image classification tasks on the AMT platform. We chose image classification, instead of other common HIT categories such as audio transcription, key point identification, or text responses because the correct answers tend to be unequivocal. To identify the treatment effect, our main approach is to regress task level performance on the assigned pay rate, controlling for other pre-treatment covariates for better precision. The resultant coefficient of assigned pay rate should be an estimate of the average pay rate effect on task level accuracy. We will walk through the motivations, designs, protocols and models for the two designs in the following sections.

4 Experimental Design and Protocol

Our experiment connect the AMT platform HIT work flow with the Qualtrics platform survey work flow. The AMT platform allows us, as a requestor to post HITs of different treatment pay rates and availabilities. Once a turker select our HIT out of a list of other HITs from other requestors based on our pay rate and description(printed below), the turker will be directed to our Qualtrics survey through a web link. Once all the survey questions are completed and the Qualtrics survey ends, the turker will submit their identification number of the AMT platform again. Once all the available HITs for a particular posting are claimed, completed and submitted by turkers, both the AMT posting and Qualtrics survey are terminated. Finally, we download data from both platforms, conduct statistical analyses and reward turkers who score higher than a pre-determined accuracy threshold.

> *Title : Multiple-Choice Task Description: This is a 50-question multiple choice task Keywords: survey, multiple-choice*
>
> *Reward per assignment: 0.1 Time allotted: 20min* (If this is too long, turkers may think this is a very hard task)
>
> *We need help with this multiple-choice task, which will provide us examples to train a computation model. The survey consist of several demographic questions, followed by 50 multiple choice questions. You don't need any prior experience or knowledge to complete this task. Select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.*
>
> *Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.*
>
> (And below this they see the survey link and the box to enter the code.)

The Qualtrics survey begins by prompting for the turker's identification number and a block of 5 forced-response, multiple-choice questions to probe the turker's aptitude for dog-breed classification. Up until this point, the turker has no knowledge that this is an image classification task, nor relevancy to dog breeds. The turker cannot revisit this question block later. Below, the questions are listed with their answer choices and intended purposes:

Number	Question	Answer Choices	Intended purpose
1	What portion of your friends own pets?	a lot less than half, around half, a lot more than half	Does the turker live in a dog owning culture?
2	Please rank your preferences to work with the following media.	audio,text,images,other	Does the turker have a strong preference for image classification?
3	Have you ever lived with any dogs in your household? If not, have you ever planned to own a dog?	Yes, Maybe, No	Foes the turker pay attention to dog breeds at all?
4	On average, how many tasks on Amazon Mechanical Turk do you complete every week?	0 to 10, 11 to 20, 21 to 30, 31 to 40, 40 or more	How much does the turker depend on Amt as a source of income?

Number	Question	Answer Choices	Intended purpose
5	Do you use Linkedin? (no need to provide any links)	Yes, No, Never heard of Linkedin	Does the turker has college or higher education? Does the turker take career development seriously?

Then, an external web link for dog breed references is provided, followed by 48 classification questions in multiple-choice format on the Qualtrics form. For the design of these classification questions, we chose eight dog breeds with a balance in size and hairy density (footnote). Even numbered questions are harder and odd numbered questions are easier. A pilot was used to identify and filter out questions which all turkers scored correctly or incorrectly. The order of questions is randomized and show a balance of even and odd numbered questions even when we split the question set into three batches. Screener questions of cat images are mixed-in to help us identify those turkers who were not paying much attention to the task. All images come from the Stanford Dogs Dataset (footnote).

However, this simple mechanism poses a threat on the unbiasedness of our estimate. Since turkers self-select into HITs, HITs of different pay rates tend to attract different kinds of turkers. From our prior internet research, turkers tend to be strategic with how their time and expectation matches with pay rate, allotted time and nature of the posted HITs. If we randomize treatment pay rate at the posting level, we would be comparing groups of turkers with different attributes. Therefore, we came up with two experiment designs which branches from the basic mechanism described above.

Design 1 is a traditional between-subject design, we define its unit of analysis as a HIT. Meaning, we place ourselves in the perspective of a data scientist in private industry who invest a company's money on getting human labeled examples for machine learning purposes. Our primary goal is to estimate how much more the company should spend on the AMT platform in order to get more accurate labeled training examples. With this motivation, we do not care about comparability of turker attributes, rather the returned accuracy per HIT as a result of different company spending. As such, selection bias and attrition from turkers are not concerns. (How do we randomize?)

Design 2 is a stepped-wedge design, we define its unit of analysis as a turker. Meaning, we place ourselves in the perspective of an economist, who studies the effect of incentives on labor productivity. Our primary goal is to estimate how increments of payment motivates a turker to perform better. With this motivation, unlike design 1, we care about comparability of turker attributes and want to ensure that treatment groups on average comprise of turkers of similar motivations and backgrounds. As such, selection bias and attrition are large concerns. In the following paragraphs we walk through each design in terms of level of randomization, treatments and execution protocol.

In design 1, we randomize at the level of HIT postings. Over two weekends in November 2017, we released eight HIT postings, that is two for each of the four different pay rates. It is a traditional between subject design with clustered randomization. Since it would not be possible randomly post HITs one at a time, we posted them in batches of 100 HITs, each batch correspond to a single pay rate. We manually shuffle the order of postings to minimize order and time of day effects. The four pay rates are chosen according to the typical minimum and maximum of other HITs alike. Time frame for the eight postings do not overlap with each other. Design 1 details are summarized below:

Experiment Schedule for Design 1

Order	Date	Time Frame	Treatment (Pay Rate)
Pilot	Oct 28, 2017 (Saturday)	Morning	$0.10
Pilot	Oct 29, 2017 (Sunday)	Afternoon	$0.25
1	Nov 11, 2017 (Saturday)	Morning	$0.10
1	Nov 11, 2017 (Saturday)	Afternoon	$0.55
1	Nov 12, 2017 (Sunday)	Morning	$0.25

Order	Date	Time Frame	Treatment (Pay Rate)
1	Nov 12, 2017 (Sunday)	Afternoon	$0.40
2	Nov 18, 2017 (Saturday)	Morning	$0.40
2	Nov 18, 2017 (Saturday)	Afternoon	$0.25
2	Nov 19, 2017 (Sunday)	Morning	$0.55
2	Nov 19, 2017 (Sunday)	Afternoon	$0.10

Design 1 Notation: Between Subject Design

R T(0.10) O

R T(0.25) O

R T(0.40) O

R T(0.55) O

In design 2, we randomize at the level of turkers instead of postings. On November 26 2017 (Sunday), we released one HIT posting of 240 available HITs and baseline rate of $0.22. It is a typical stepped-wedge design with randomization at the turkers level. Turkers would sign up for the HIT for the same baseline rate, and then randomized with equal probability into one of four treatment groups after they submitted their identification number and aptitude question answers on the Qualtrics survey form. The treatment group differs by the amount of surprise bonuses (up until this point the turker has no knowledge that this task may come with any bonuses). Here, the 48 dog breed classification questions from design 1 are split into three sesssions of 16 questions. The overall question order is the same as that in design 1, and the three sessions share a balance of difficulty and dog breeds. Each session is associated with a bonus assignment condition of either $0.10 or nothing with no mention of bonus condition at all. We chose the baserate as $0.22 rather than $0.10 to minimize attrition and set the total available HITs to be 240 so to stay within experiment budget. Design 1 details are summarized below:

Experiment Schedule for Design 2

Name	Date	Time Frame	Base pay rate	Treatments (bonuses)
Pilot	Nov 23, 2017 (Thursday)	All day	$0.10	$0.00, $0.05, $0.10, $0.15
Main	Nov 26, 2017 (Sunday)	All day	$0.22	$0.00, $0.10, $0.20, $0.30

Design 2 Notation: Stepped-Wedge Design

R C(0.00) O C(0.00) O C(0.00) O

R C(0.00) O C(0.00) O T(0.10) O

R C(0.00) O T(0.10) O T(0.10) O

R T(0.10) O T(0.10) O T(0.10) O

For both experiments, we took specific cautions in our execution protocol. Our pilots results indicated that although the pool size of Amazon turkers is in the order of hundred thousands, several turkers managed to find and submit our HIT for again but for a different pay rate. Additionally, some turkers may check out

the HIT, go through the covariate questions, take a look at the dog breed classification questions, leave the HIT at one pay rate and sign up again for another higher pay rate. Therefore, in both designs, we assign turkers with "qualifications" – labels with which we filter out turkers who have completed our HITs from the pool of turkers who may continue to see our following postings. We also keep a continuously updated list of identification numbers of those turkers who attrited, so to conditionally block them from accessing our Qualtrics survey. Because multiple attempts or preview of the same task under different treatment conditions would have carry-over or spill-over effects on the outcome, we felt that these cautions were necessary. On the other hand, differential attrition of turkers, although not specifically analyzed in design 1 (since our unit of analysis is defined as the returned HIT rather than the turker), was conspicuous in the data. To mitigate the problem that turkers who started in lower pay rate postings tend to attrite more than those who started in higher pay rate postings, we raise the base rate in design 2, in which turkers are our unit of analysis, from \$0.10 to \$0.22. The results section will give describe attrition data in detail.

A note on supplementary files: A clean, well organized Github repo is available for this project. It includes raw, intermediate and transformed data, along with data transformation, statistical and project management codes. Because of the length of extensiveness of these R functions, we decided to print such functions as a higher level of abstraction and maintain a natural presentation flow of the project. For project evaluation's sake, please refer to the .rmd file for our the hidden snippets, and our Github repo for any other materials. (Available upon request)

---

5 Design 1 Results

5.1 Pilot Study (Design 1)

By running a pilot study for design 1, we tested the experiment protocol, identified problems in our AMT and Qualtrics workflow and conducted a power analysis on the collected data. During the last weekend of Oct 2017, we published two non-overlapping HIT postings each with 50 HITs available. One at the treatment pay rate of \$0.10 on Saturday and the other at the treatment pay rate of \$0.25 on Sunday. We tried to minimize differences in launching conditions for the two postings so to ensure comparability.

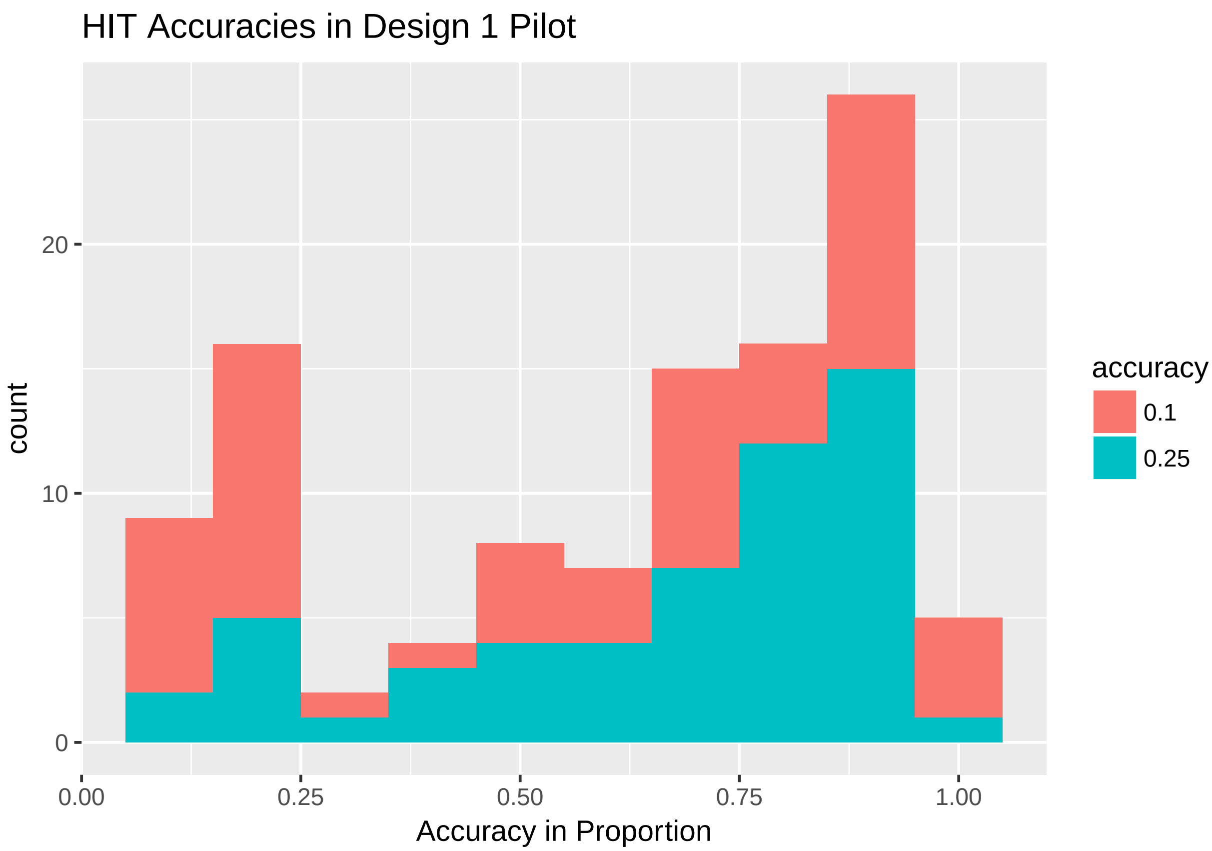
5.1.1 Data (Design 1 Pilot)

The below table shows a summary of the two pilot postings and corresponding average accuracies. In comparison, we can see that the posting that paid more returned a higher average accuracy and completed faster than the lower paying posting. Note that the number of returned HITs in each posting is higher than 50 because some turkers may submit the HIT before the Qualtrics survey terminates but after the AMT posting terminates.

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Pilot 1	\$0.10	54	2h 30min	5.317min	0.559	0.320
Pilot 2	\$0.25	54	1h 20min	5.837min	0.673	0.246

The overall accuracy distribution is bimodal. One mode occurs between [0.15,0.20] and the other occurs between [0.85,0.95]. The distribution of HITs listed for \$0.10 bias towards the first mode, while that for \$0.25 bias towards the second mode. This is inline with our expectation that turkers are either accomplish with determination or care little (given eight choices for each question, making random choices would yield an accuracy of 0.125 in expectation). The fact that this accuracy distribution is non-normal cautioned us against reliance on OSL asymptotics for standard error estimation. While a larger sample size can increase

this reliability, we nevertheless plan to include randomization inference on top of the t-statistic from OLS.



5.1.2 Covariate Balance (Design 1 Pilot)

Of the 5 aptitude questions we asked of our turkers, we believe that responses to question 1, 2 and 3 do not depend on the treatment assignment, since turkers have no knowledge of the task being related to image classification nor dog breeds until this point of the survey. In contrast, responses to question 4 and 5 probes the turkers' income and education level, so they are prone to selection bias associated with the posted HIT payrate. Therefore, we assume responses to the question 1, 2 and 3 are useful controls while the other two are bad controls. To conduct a covariate balance check, we regress responses to question 1, 2 and 3 on the treatment variable. The regression table summarizes that treatment fails to predict any of the answers in a statistically significant way. Our covariate balance check has passed.

```
##
## Covariate Balance Check
## ====================================================================================
##                                                   Dependent variable:
##                                   -------------------------------------------------------
##                                    CQ1_1   CQ1_2   CQ1_3   CQ2_3    CQ3_1    CQ3_2   CQ3_3
##                                     (1)     (2)     (3)     (4)      (5)      (6)     (7)
## ------------------------------------------------------------------------------------
## treatment                         -0.370   0.370  -0.000   -1.481   -0.370   0.370  -0.000
##                                   (0.534) (0.647) (0.624) (0.985)  (0.534)  (0.534) (0.534)
##
## Constant                          0.278** 0.370** 0.352** 2.148*** 0.852***  0.056   0.093
##                                   (0.105) (0.123) (0.119) (0.195)  (0.098)  (0.098) (0.098)
```

```
##
## ---------------------------------------------------------------------------------------------
## Observations                        108     108     108     108      108      108     108
## R2                                 0.005   0.003   0.000   0.021    0.005    0.007   0.000
## Adjusted R2                       -0.005  -0.006  -0.009   0.012   -0.005   -0.002  -0.009
## Residual Std. Error (df = 106)     0.412   0.500   0.482   0.761    0.412    0.327   0.293
## F Statistic (df = 1; 106)          0.490   0.334   0.000   2.304    0.490    0.778   0.000
## =============================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

CQ1_1 indicates if the turker has a lot less than half of friends who own pets, CQ1_2 indicates if the turker has around half of friends who own pets, CQ1_3 indicates if the turker has a lot more than half of friends who own pets, CQ2_3 ranks turkers' preference to work with images, CQ3_1 indicates that the turker has lived with or planned to own a dog, CQ3_2 indicates that the turker may have planned to own a dog, CQ3_3 indicates that the turker has never lived with or planned to own a dog

5.1.3 ATE Estimation (Design 1 Pilot)

We performed a basic power analysis on our pilot data, so we could get an initial feel of the results we would be getting. First, we conducted a Levene test, which showed us that the variances of the $0.10 outcomes and the $0.25 outcomes are similar. From there, we ran a two-sample independent t-test, as well as a simple and a full regression with robust standard error.

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 24  0.7465 0.7658
##       29
```

```
##
##  Two Sample t-test
##
## data:  worker_perf_pilot_0.10$accuracy and worker_perf_pilot_0.25$accuracy
## t = -2.0644, df = 106, p-value = 0.04142
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.222176098 -0.004490568
## sample estimates:
## mean of x mean of y
## 0.5592593 0.6725926
```

```
##
## Covariate Balance Check
## ==================================================================
##                                    Dependent variable:
##                          -----------------------------------------
##                                          accuracy
##                               simple              full
##                                 (1)                (2)
## ------------------------------------------------------------------
## treatment                     0.756*              0.651
##                              (0.369)             (0.359)
##
## CQ1a lot more than half                           0.146
##                                                  (0.081)
```

```
##
## CQ1around half                                          0.002
##                                                        (0.082)
##
## CQ2_3                                                 -0.072*
##                                                        (0.036)
##
## CQ3No                                                  0.119
##                                                        (0.146)
##
## CQ3Yes                                                 0.126
##                                                        (0.127)
##
## Constant                          0.484***             0.472**
##                                   (0.077)              (0.168)
##
## ------------------------------------------------------------------
## Observations                        108                   108
## R2                                 0.039                 0.154
## Adjusted R2                        0.030                 0.103
## Residual Std. Error          0.285 (df = 106)     0.274 (df = 101)
## F Statistic                4.262* (df = 1; 106) 3.054** (df = 6; 101)
## ==================================================================
## Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

6 Design 2 Results

- pilot
- main

7 Future Lines of investigation

8 Conclusion and Bibliography