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, 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.

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?
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	Treatment (Pay Rate)	Available HITs
1	Nov 11, 2017 (Saturday)	\$0.10	100
1	Nov 11, 2017 (Saturday)	\$0.55	100
1	Nov 12, 2017 (Sunday)	\$0.25	100
1	Nov 12, 2017 (Sunday)	\$0.40	100
2	Nov 18, 2017 (Saturday)	\$0.40	100
2	Nov 18, 2017 (Saturday)	\$0.25	100
2	Nov 19, 2017 (Sunday)	\$0.55	100
2	Nov 19, 2017 (Sunday)	\$0.10	100

Design 1 Notation: Between Subject Design

R X(0.10) O

R X(0.25) O

R X(0.40) O

R X(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

Date	Base pay rate	Treatments (bonuses)	Available HITs
Nov 26, 2017 (Sunday)	$0.22	$0.00, $0.10, $0.20, $0.30	240
Nov 26, 2017 (Sunday)	$0.22	$0.00, $0.10, $0.20, $0.30	240

Design 2 Notation: Stepped-Wedge Design

R X(0.00) O X(0.00) O X(0.00) O

R X(0.00) O X(0.00) O X(0.10) O

R X(0.00) O X(0.10) O X(0.10) O

R X(0.10) O X(0.10) O X(0.10) O

- Protocol

Add pilots to schedule For both experiments, we took specific cautions in our execution protocol, to prevent/account.......

5 Results – Design 1

6 Results – Design 2

7 Future Lines of investigation

8 Conclusion and Bibliography