

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.

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, or a turker. Similarly, in each of the two experimental designs, a total of 50 image classifications questions 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 motivation, design, protocol and models for the two designs in the next section.

4 Experimental Design and Protocol

- Challenge to answer research question – why two designs
- unit of analysis
- advantages and problems of design 1
- advantages and problems of design 2
- Design 1
- Treatment groups
- Experiment Schedule
- Protocol
- Design 2
- Treatment groups
- Time and Date
- Protocol

5 Results – Design 1

6 Results – Design 2

7 Future Lines of investigation

8 Conclusion and Bibliography