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

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?
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.

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 randomization at the cluster level. 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
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.

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.

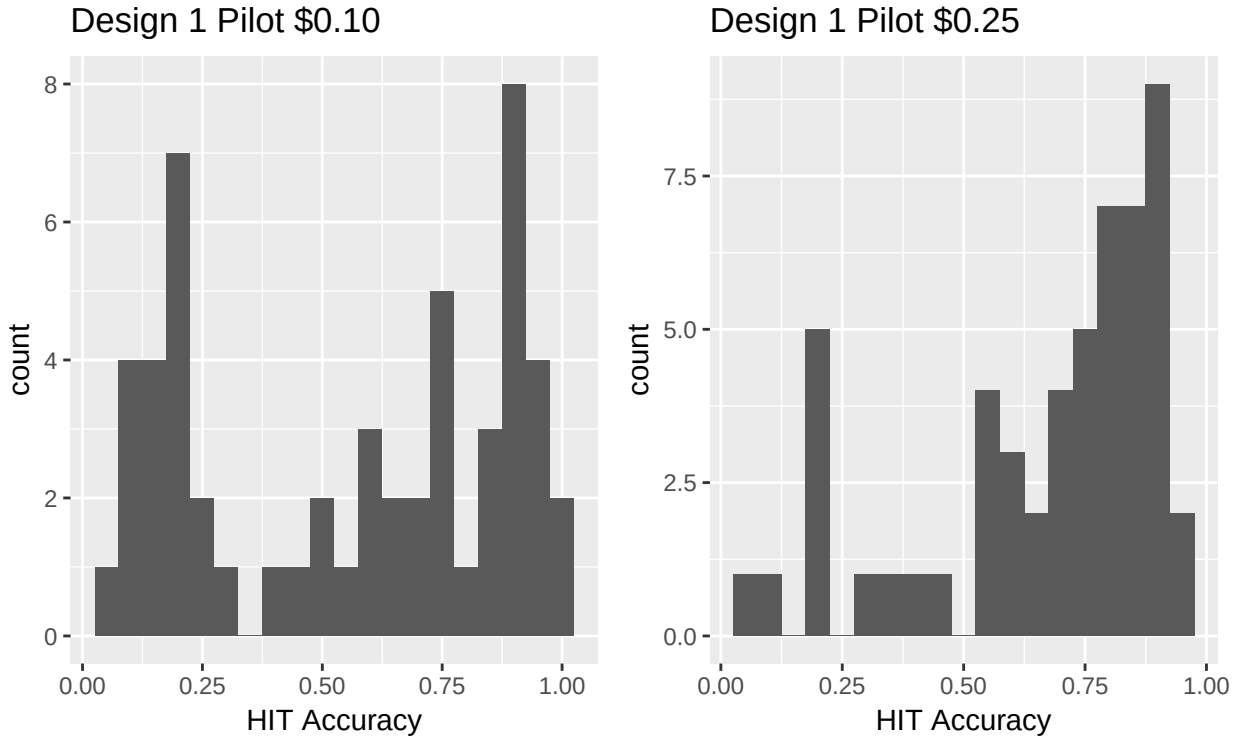
5.1.1 Data (Design 1 Pilot)

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 Design 1 Pilot
## ================================================================================================
##                                                 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 Treatment Effect 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 significantly different. From there, we ran a two-sample independent Welch's t-test, as well as a simple and a full regression with robust standard error. Below are the two models we use to estimate ATE in the pilot:

(simple) $accuracy = \theta_0 + \theta_1 * treatment$

(full) $accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3$

The two-sample independent t-test results, which is equivalent to OLS results with only one variable, shows marginal significance for the treatment effect of 0.756 higher accuracy per dollar spent. The full-regression shows no statistical significance (p-val ~ 0.07) for the treatment effect of 0.651 higher accuracy per dollar spent. These results lead us to believe that there might be statistical significance in the main experiment we planned. Therefore, for in the main experiment for design 1, we decided to raise sample size for each posting from 50 to 100, and further expand our treatments to include $0.40 and $0.55.

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  8.2433 0.00494 **
##       106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Welch Two Sample t-test
##
## data:  worker_perf_pilot_0.10$accuracy and worker_perf_pilot_0.25$accuracy
## t = -2.0644, df = 99.394, p-value = 0.04158
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.222259674 -0.004406993
## sample estimates:
## mean of x mean of y
## 0.5592593 0.6725926
```

```
##
## Design 1 Pilot ATE Estimation
## ========================================================================
##                                    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
```

```
## Simple Model with No Covariates:
```

```
## estimated average causal effect =  0.7555556
## robust standard error =  0.3694305
```

```
## 95% confidence interval =  0.02312365 1.487987
## p-value =  0.04331308

##
##
## Full Model with Covariates:

## estimated average causal effect =  0.6507679
## robust standard error =  0.3587494
## 95% confidence interval =  -0.06089447 1.36243
## p-value =  0.07264833
```

5.2 Main experiment (Design 1)

We ran the main Design 1 (D1) experiment over the third of fourth weekends of November. In each weekend we published four postings each of 100 available HITs, and cover all four treatment pay rates. On each day, we publish one posting in the morning, and another one in the afternoon. The two postings don't overlap, and we take aforementioned precautions to make sure turkers who had done or seen the task once couldn't do it again. The only difference between the two weekends were the order in which we released the prices (order1 is 0.10, 0.55, 0.25, 0.40; order 2 is 0.40, 0.25, 0.55, 0.10). This allows us to check whether the order of the prices actually makes a difference.

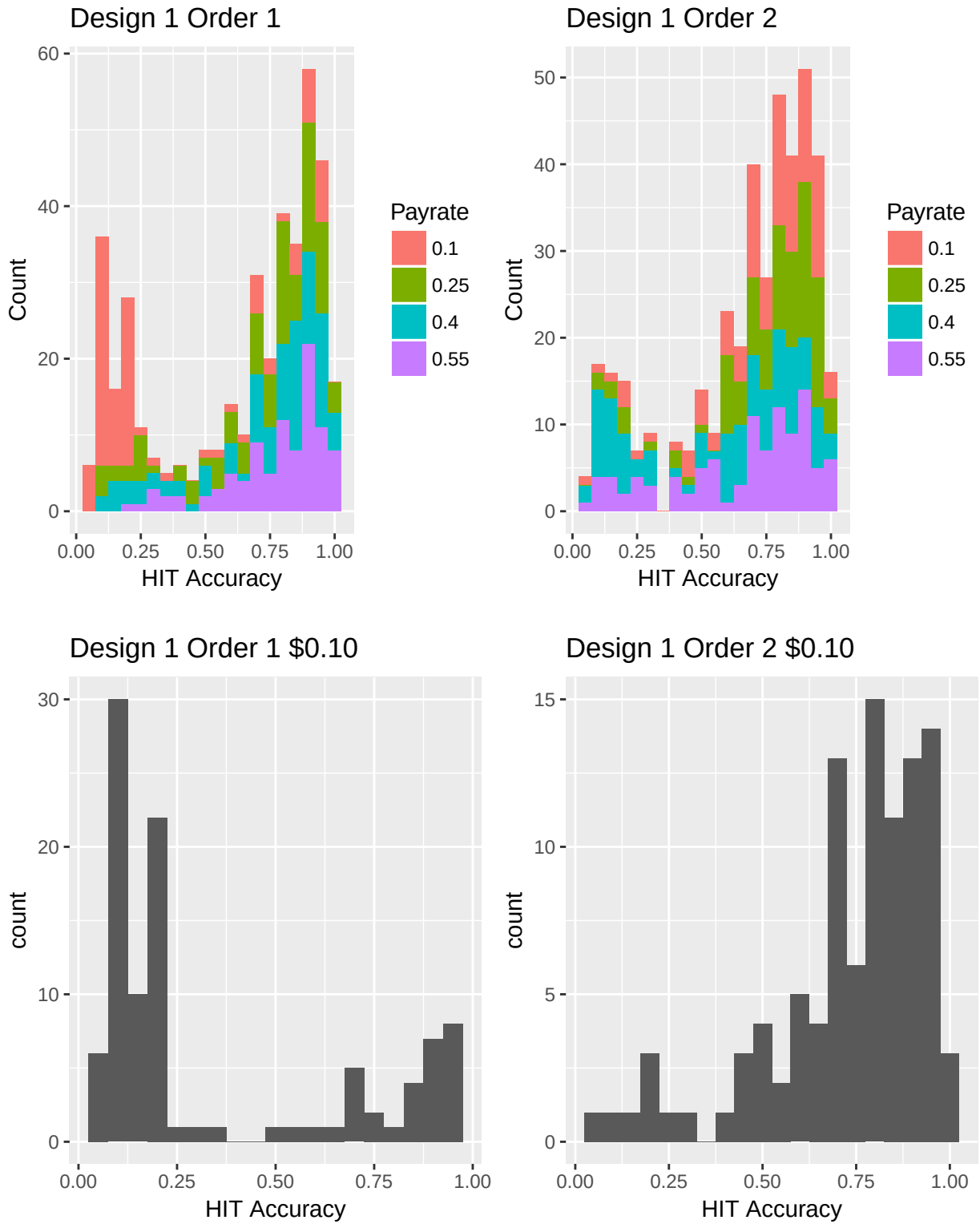
5.2.1 Data (Design 1 Main)

Summary of Design 1 data is provided in the table below. Similar to what we saw in the pilot, available HITs of postings of higher pay rates are claimed much sooner, even though completion time per HIT stayed roughly the same. It is likely due to competition – turkers want to avoid missing out well paid HITs before they disappear, while poorly paid HITs are less in demand. One can interpret this difference in time for all HITs to be calimed to be a good indicator that the chosen treatment pay rates are varied enough to stimulate differential reactions from the turker workforce. Even so, selection bias cannot be ruled out. For example, only the faster workers, or the ones with the best internet connection, will be able to do the higher paying HITs.

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Order 1	$0.10	102	11h 15min	0.912min	0.345	0.327
Order 1	$0.25	103	3h 15min	0.922min	0.703	0.253
Order 1	$0.40	102	0h 40min	0.866min	0.716	0.256
Order 1	$0.55	98	0h 40min	0.666min	0.774	0.192
Order 2	$0.10	102	15h 20min	0.79min	0.734	0.216
Order 2	$0.25	102	2h 40min	0.832min	0.752	0.217
Order 2	$0.40	105	2h 20min	0.855min	0.562	0.306
Order 2	$0.55	103	3h 10min	0.912min	0.656	0.265

In terms of average accuracy, the distributions of HIT accuracy are again bimodal and suggest the need for randomization inference for causal effect estimation. Order 1 data shows similar distribution as the pilot, with low paying HITs biased towards the lower mode and high paying HITs biased towards the higher mode. Although order 2 data has a similar overall distribution, differential distribution among HITs of different pay rates is not observed anymore. From the earlier summary table, we observe an abnormality in the accuracy for the two $0.10 postings – average accuracy of the first one is roughly one standard deviation lower than the second one. Therefore, the $0.10 HITs collected during the first weekend, like the pilot, has a noticeable right skew while the $0.10 HITs collected during the second weekend follows the overall trend of all HITs. Such a conspicuous difference between outcomes from similar treatments suggests that there is either an

order effect, Saturday versus Sunday effect, morning vs afternoon effect, or an effect coming from outside of the experiment system.



5.2.2 Covariate Balance (Design 1 Main)

The covariate balance check on the combined D1 data doesn't show any statistically significant results for the answers to our covariate questions. The same check on the order levels reveals similar results. This gives us confidence that the covariates are uncorrelated with treatment assignment and suitable as controls.

```
##
## Covariate Balance Check Design 1 Overall
## ================================================================================================
##                                                Dependent variable:
##                                ---------------------------------------------------------------
##                                 CQ1_1    CQ1_2    CQ1_3    CQ2_3    CQ3_1    CQ3_2    CQ3_3
##                                  (1)      (2)      (3)      (4)      (5)      (6)      (7)
## ------------------------------------------------------------------------------------------------
## treatment                       0.008    0.001   -0.009    0.064   -0.010    0.051   -0.041
##
##
## Constant                        0.199   0.464*** 0.337** 1.775*** 0.842***  0.073    0.086
##                                (0.105) (0.123)  (0.119) (0.195)  (0.098)  (0.098)  (0.098)
##
## ------------------------------------------------------------------------------------------------
## Observations                     817      817      817      817      817      817      817
## R2                              0.00001 0.00000  0.00001  0.0002  0.00002   0.001    0.001
## Adjusted R2                     -0.001   -0.001   -0.001   -0.001   -0.001  -0.0003  -0.001
## Residual Std. Error (df = 815)  0.402    0.499    0.472    0.757    0.369    0.285    0.259
## F Statistic (df = 1; 815)       0.009    0.0001   0.008    0.165    0.018    0.738    0.573
## ================================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
##
## Covariate Balance Check Design 1 Order 1
## ================================================================================================
##                                                Dependent variable:
##                                ---------------------------------------------------------------
##                                 CQ1_1    CQ1_2    CQ1_3    CQ2_3    CQ3_1    CQ3_2    CQ3_3
##                                  (1)      (2)      (3)      (4)      (5)      (6)      (7)
## ------------------------------------------------------------------------------------------------
## treatment                       -0.055   -0.100    0.155             0.094   -0.013   -0.082
##
##
## treatment                                          0.184
##
##
## Constant                        0.223*  0.494*** 0.283*  1.686*** 0.831***  0.086    0.083
##                                (0.105) (0.123)  (0.119) (0.195)  (0.098)  (0.098)  (0.098)
##
## ------------------------------------------------------------------------------------------------
## Observations                     405      405      405      405      405      405      405
## R2                               0.001    0.001    0.003    0.002    0.002   0.0001    0.003
## Adjusted R2                     -0.002   -0.001    0.001   -0.001  -0.0004  -0.002    0.001
## Residual Std. Error (df = 403)   0.405    0.499    0.472    0.736    0.346    0.274    0.232
## F Statistic (df = 1; 403)        0.212    0.450    1.220    0.706    0.842    0.025    1.398
## ================================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

```
##
## Covariate Balance Check Design 1 Order 2
## ================================================================================================
##                                                  Dependent variable:
##                                 ---------------------------------------------------------------
##                                   CQ1_1     CQ1_2     CQ1_3     CQ2_3     CQ3_1     CQ3_2     CQ3_3
##                                    (1)       (2)       (3)       (4)       (5)       (6)       (7)
## ------------------------------------------------------------------------------------------------
## treatment                         0.070     0.099    -0.169               -0.110     0.113    -0.003
##
##
## treatment                                                    -0.059
##
##
## Constant                          0.176   0.434***  0.390**  1.864***  0.851***   0.060     0.088
##                                  (0.105)  (0.123)   (0.119)  (0.195)   (0.098)   (0.098)   (0.098)
##
## ------------------------------------------------------------------------------------------------
## Observations                       412       412       412       412       412       412       412
## R2                                0.001     0.001     0.004    0.0002     0.002     0.004   0.00000
## Adjusted R2                      -0.002    -0.001     0.001    -0.002   -0.0002     0.002    -0.002
## Residual Std. Error (df = 410)    0.400     0.500     0.472     0.775     0.388     0.296     0.283
## F Statistic (df = 1; 410)         0.356     0.451     1.478     0.067     0.922     1.685     0.002
## ================================================================================================
## Note:                                                                *p<0.05; **p<0.01; ***p<0.001
```

Definition of covariate variables remains unchanged in all experiments. Please reference the design 1 pilot version.

5.2.3 Treatment Effect Estimation (Design 1 Main)

5.2.3.1 T-tests (Design 1 Main)

We start off by running independent two-sample t-tests, doing a pairwise comparison between the accuracies for all treatments. We used either a normal t-test or a Welch two sample t-test, depending on similarity in variances between treatment group outcomes. Because we are performing multiple comparisons with 18 tests based on the same data source and the same research question, we use Bonferroni correction to set our significance cut-off as 0.002777778.

When data from the two orders are pooled together, accuracies from \$0.10 HITs are significantly lower than higher pay rate HITs. The comparison between \$0.25 and \$0.40 HITs is significant but in opposite direction to the research hypothesis. The comparisons between \$0.25 or \$0.40 HITs against \$0.55 show no significance. With data from order 1, same results are only observed in the comparisons involving \$0.10 HITs. With data from order 2, only the \$0.10 vs \$0.40, and \$0.25 vs \$0.40 comparisons are significant. Overall, the \$0.10 s \$0.40 comparison is the only consistent one across both orders. However, their p values are quite different and their mean difference have opposite signs. Simply looking at the t-tests results, the effect significance is quite inconclusive.

```
##
## Independent two-sample t-test results for Design 1 treatment group comparisons
```

Data	Comparison	t.stat	N1	N2	df	Mean1	Mean2	up.CI	low.CI	p.val	significance
Pooled	\$0.10 vs \$0.25	-6.511	204	205	363.048	0.539	0.728	-0.245	-0.131	0.000000	1
Pooled	\$0.10 vs \$0.40	-3.158	204	207	398.721	0.539	0.638	-0.160	-0.037	0.001708	1
Pooled	\$0.10 vs \$0.55	-5.990	204	201	365.544	0.539	0.714	-0.231	-0.117	0.000000	1

Data	Comparison	t.stat	N1	N2	df	Mean1	Mean2	up.CI	low.CI	p.val	significance
Pooled	$0.25 vs $0.40	3.420	205	207	394.513	0.728	0.638	0.038	0.141	0.000692	1
Pooled	$0.25 vs $0.55	0.588	205	201	404.000	0.728	0.714	-0.033	0.060	0.556955	0
Pooled	$0.40 vs $0.55	-2.866	207	201	394.756	0.638	0.714	-0.128	-0.024	0.004375	0
Order 1	$0.10 vs $0.25	-8.768	102	103	203.000	0.345	0.703	-0.439	-0.278	0.000000	1
Order 1	$0.10 vs $0.40	-9.022	102	102	202.000	0.345	0.716	-0.453	-0.290	0.000000	1
Order 1	$0.10 vs $0.55	-11.385	102	98	164.208	0.345	0.774	-0.504	-0.355	0.000000	1
Order 1	$0.25 vs $0.40	-0.370	103	102	203.000	0.703	0.716	-0.083	0.057	0.711883	0
Order 1	$0.25 vs $0.55	-2.250	103	98	199.000	0.703	0.774	-0.134	-0.009	0.025528	0
Order 1	$0.40 vs $0.55	-1.819	102	98	198.000	0.716	0.774	-0.122	0.005	0.070381	0
Order 2	$0.10 vs $0.25	-0.602	102	102	202.000	0.734	0.752	-0.078	0.041	0.547897	0
Order 2	$0.10 vs $0.40	4.688	102	105	187.305	0.734	0.562	0.100	0.244	0.000005	1
Order 2	$0.10 vs $0.55	2.324	102	103	195.714	0.734	0.656	0.012	0.145	0.021179	0
Order 2	$0.25 vs $0.40	5.176	102	105	187.794	0.752	0.562	0.118	0.263	0.000001	1
Order 2	$0.25 vs $0.55	2.858	102	103	196.074	0.752	0.656	0.030	0.163	0.004723	0
Order 2	$0.40 vs $0.55	-2.359	105	103	206.000	0.562	0.656	-0.172	-0.015	0.019284	0

We can also run a t-test to compare the outcomes of the two different orders. Doing this gives us a p-value of 0.038, suggesting that the order or weekend with which we published the postings might have an effect on accuracy. This is consistent with our earlier observation that $0.10 HITs outcomes from the two weekeneds are very different. Yet, given that multiple sub-tests returned significant results, we take this as an indiciator that there may exists some relationship between payment and accuracy, and we explore this using regression in the next step.

```
##
##  Welch Two Sample t-test
##
## data:  d1 and d2
## t = -2.0698, df = 789.3, p-value = 0.0388
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.081561097 -0.002160586
## sample estimates:
## mean of x mean of y
## 0.6333333 0.6751942
```

5.2.3.2 Linear regression (Design 1 Main)

We construct six linear regression models on the dataset, their specifications are provided in the formulas below. OLS estimates for model 1 provides an estimate for the average treatment effect on accuracy. Model 2 controls for covariates to improve the precision of our average treatment effect. Model 3 further controls for the order and additionally detect any order 1 or first weekend effects. Model 4 further includes an interaction term between treatment and order to look for heterogeneous treatment effects – whether receiving treatment in order 1 (first weekend) is better or worse than receiving the same treatment in order 2 (second weekend). Model 5 further includes a second-order term to see if the data fits better, to give us further insight into the statistical relationship between remuneration and worker performance. Note that this model is not supported by prior hypothesis or intuitions of the statistical relationship between outcome and treatment. We include this simply out of curiosity. Model 6 drops the second order term in model 5 but further include interaction between treatment and covariates to detect any heterogeneous treatment effect. Note that model 1, 2, 3, 4 and 6 assume that the relationship between accuracy and treatment is linear (first order).

(1) $accuracy = \theta_0 + \theta_1 * treatment$

(2) $accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3$

(3) $aaccuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$

(4) $accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$

(5) $accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1 + \theta_7 * treatment^2$

(6) $accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1 + \theta_8 * treatment^2 * cq1 + +\theta_9 * treatment^2 * cq1 + +\theta_{10} * treatment^2 * cq1$

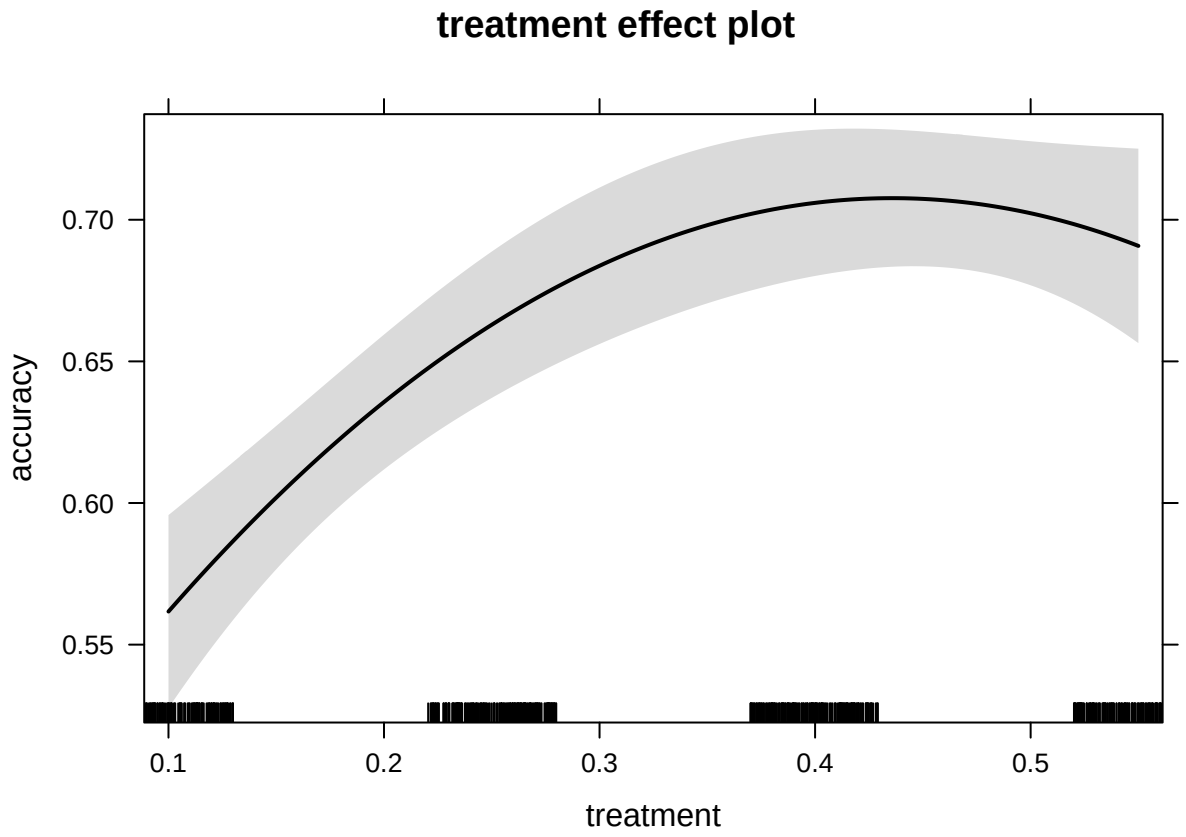
```
##
##  Model 1: simple: ATE
## estimated average causal effect =  0.288234
## Clustered standard errors = 0.3249669
## .95 CI with clustered SE = [ -0.3496367 0.9261047 ]
## p-value =  0.3753599

##
##  Model 2: add covariates: ATE
## estimated average causal effect =  0.287949
## Clustered standard errors = 0.3175239
## .95 CI with clustered SE = [ -0.3353178 0.9112158 ]
## p-value =  0.3647518

##
##  Model 3: add Order indicator: ATE
## estimated average causal effect =  0.2864514
## Clustered standard errors = 0.2969405
## .95 CI with clustered SE = [ -0.2964133 0.869316 ]
## p-value =  0.3349954

##
##  Model 4: add interaction term: ATE for Order 1
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect = $ 0.8423

## Clustered standard errors = 0.2517

## .95 CI with clustered SE = [ 0.348237 1.336363 ]

## p-value = 0.000857

##
##  Model 4: add interaction term: ATE for Order 2
## estimated average causal effect =  -0.2566776
## Clustered standard errors = 0.1028595
## .95 CI with clustered SE = [ -0.458581 -0.05477418 ]
## p-value =  0.01277931

##
##  Model 5: add second order term: ATE for Order 1
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect = $ 0.3873

## Clustered standard errors = 0.4203

## .95 CI with clustered SE = [ -0.4377102 1.21231 ]

## p-value = 0.357
```

```
##
##  Model 5: add second order term: ATE for Order 2
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect = $ -0.7099

## Clustered standard errors = 0.4201

## .95 CI with clustered SE = [ -1.534518 0.1147176 ]

## p-value = 0.0914

##
##  Model 6: add treatment-covariate interaction: ATE for Order 1
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect = $ 0.857

## Clustered standard errors = 0.2752

## .95 CI with clustered SE = [ 0.3168037 1.397196 ]

## p-value = 0.00191

##
##  Model 6: add treatment-covariate interaction: ATE for Order 2
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect = $ -0.2324

## Clustered standard errors = 0.1138

## .95 CI with clustered SE = [ -0.4557806 -0.009019405 ]

## p-value = 0.0414

##
## Model 4 inference for coefficient covariate Order 1:

##      Estimate    Std. Error       t value      Pr(>|t|)
## -0.4057450730  0.1191874652 -3.4042596045  0.0006961302
```

Because HITs were assigned with different pay rates in groups (per posting published on AMT platform), we use clustered standard errors for inferences. Model 1, 2, 3, 5 and 6 estimated no significant treatment effects.

In model 4, estimated ATE for order 1 is significant with a point estimate of \$0.8423, with a .95 confidence interval between 0.348 and 1.3363. That is, for every additional \$0.10 spent on an HIT, we should expect an increase in accuracy between 0.0348 and 0.13363. Estimated ATE for order 2 is marginally significant with a p value of 0.01277931. But since we are performing 9 multiple inferences (3 ATEs for model 1-3, 6 ATEs for model 4-6), we apply Bonferroni correction to adjust the critical cut-off level to 0.00556. After this adjustment, the estimated ATE is not significant anymore. Notice that in model 4 the best estimate for ATE in order 1 and order 2 are off opposite signs, coefficient for order 1 is also highly significant, reinforcing our earlier observation of order or weekend effect.

In model 5, neither first order nor second order treatment regressor had a significant estimate, the result fails to suggest any treatment effect nor non-linear relationship between outcome and treatment. The point estimate, at most a best guess, suggest that the second-order term is statistically significant and negative, while the first-order term is positive. This means that higher payment has a positive effect on accuracy, but this effect diminishes as the payments become higher and higher. Note that similar to model 4, the estimated ATE for order 1 is positive but that for order 2 is negative.



In model 6, estimated ATE for order 1 and order 2, controling for the covariates at their mode or mean values, are marginally significant on their own. But once Bonferroni adjustment is applied, they lost significance. Note that similar to model 4 and 5, the estimated ATE for order 1 is positive but that for order 2 is negative.

5.2.3.3 Randomization Inference (Design 1 Main)

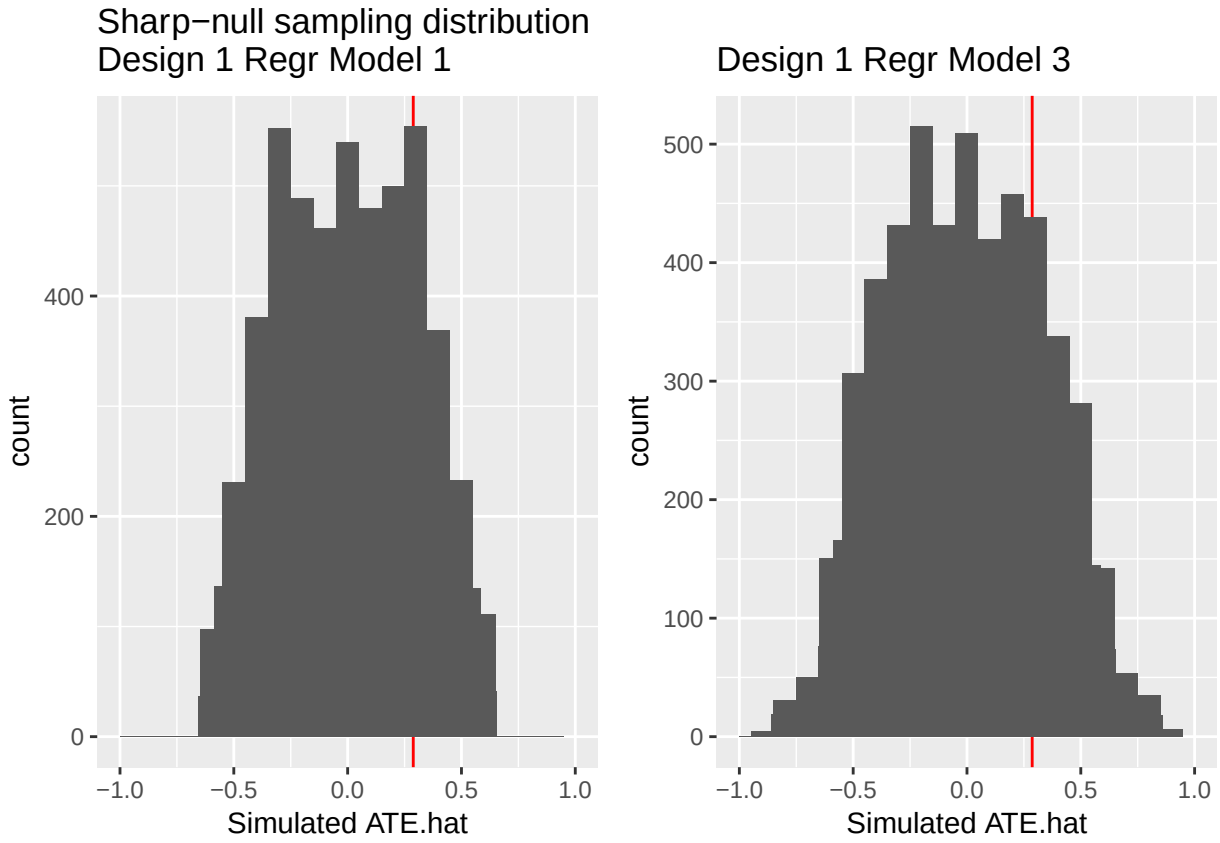
Because of the non-normality observed in our data, we also use randomization inference, to see if a more exact simulation of the sampling distribution under sharp null hypothesis would drastically change the p value of our estimate. Note that randomization is conducted at the cluster level because of the way we publish HITs. That is, we randomly assign each of the four pay rates to two of the HIT postings, each comprised somewhere between 98 to 102 collected HITs. In total, there are 2520 possible randomizations. Then, we fill out the entire potential outcome schedule assuming that the accuracy for every HIT is the same regardless of the associated pay rates. After that we estimated the ATE with two approaches. First we use simple regression model 1, regressing accuracy only on pay rate. Second, we use covariate controled regression model 3, regressing accuracy on pay rate, the covariates from responses for aptitude questions, and the order/weekend at which the HIT posting was published. The randomization inference results using either approaches shows largely insignificant effect. Compared to the t-test p-values of 0.3753599 for model 1 and 0.3349954 for model 3, the p-values randomization inference are noticeably larger at 0.4086 for model 1 and 0.4738 for model 3.

(To check validity if randomization inference code, please refer to our .rmd files)

```
## Estimated ATE from simple regression (model 1): 0.288234

## p-value under sharp-null randomization inference: 0.415

##
## Estimated ATE from covariates controled regression (model 3): 0.2864514
```

```
## p-value under sharp-null randomization inference: 0.471
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



5.2.3.4 Logistic Regression (Design 1 Main)

As an extension, although logistic regression is outside the context of w241, we acknowledge that our outcome variable – accuracy is actually bounded by [0,1] and therefore logistic regression is more appropriate with our data. Below we perform the same inference procedure with our design 1 data, but replacing linear regression with logistic regression and update our interpretation accordingly.

(1.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment$

(2.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3$

(3.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$

(4.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$

(5.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1 + \theta_7 * treatment^2$

(6.1) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1 + \theta_8 * treatment^2 * cq1 + +\theta_9 * treatment^2 * cq1 + +\theta_{10} * treatment^2 * cq1$

Similar to our linear regression models, although a logit function is now applied to our outcome variable, the coefficients that are not significant in the linear regression models are still not significant in the logistic regression models, while those that are significant in the linear regression models remain significant in the logistic regression models. Here, we focus on interpretating model 4.1 only, because only this model shows

significant ATE. Also, recall the interpretation with the linear models, ATE estimated by linear model 4 on order 1 data was the only significant effect after Bonferroni adjusted cut-off was applied.

```
##
##  Model 4: Logistic Regression : Order 1
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated coefficient of treatment = $ 3.961

## Clustered standard errors = 1.067

## .95 Wald CI with clustered SE = [ 1.866581 2.936719 ]

## p-value = 0.000205

##
##  Model 4: Logistic Regression : Order 2
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated coefficient of treatment = $ -1.226

## Clustered standard errors = 0.4767

## .95 Wald CI with clustered SE = [ -2.161716 -0.2902835 ]

## p-value = 0.0101
```

As expected, estimated coefficient

For proper multiple inference, we maintain the same critical cut-off 0.00556

In model 4, estimated ATE for order 1 is significant with a point estimate of $0.8423, with a .95 confidence interval between 0.348 and 1.3363. That is, for every additional $0.10 spent on an HIT, we should expect an increase in accuracy between 0.0348 and 0.13363. Estimated ATE for order 2 is marginally significant with a p value of 0.01277931. But since we are performing 9 multiple inferences (3 ATEs for model 1-3, 6 ATEs for model 4-6), we apply Bonferroni correction to adjust the critical cut-off level to 0.00556. After this adjustment, the estimated ATE is not significant anymore. Notice that in model 4 the best estimate for ATE in order 1 and order 2 are off opposite signs, coefficient for order 1 is also highly significant, reinforcing our earlier observation of order or weekend effect.

- Swtich outcome variable

6 Design 2 Results

6.1 Pilot Study (Design 2)

6.2 Main experiment (Design 2)

6.2.1 Data (Design 2 Main)

6.2.2 Covariate Balance (Design 2 Main)

6.2.3 Treatment Effect Estimation (Design 2 Main)

- Design 1 and 2 comparison

7 Future Lines of investigation

8 Conclusion and Bibliography

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)
