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

Legg Yeung, Stanimir Vichev, Frederic Suares

University of California, Berkeley

December 13, 2017

Abstract

In this project, we look at the relationship between reward and work quality through the different perspective of employers and economists. We use two experimental designs, one being a traditional between-subject experiment and the other being a stepped-wedge design, to analysis effects of higher pay rate on returned work quality and effects of bonuses on worker performance. We utilize the functionalities of the Amazon Mechanical Turk platform and Qualtrics survey tools to randomly assign a human intelligence task to different pay rate in the first experimental design, and a turker to different bonus conditions in the second experimental design. Both experimental setups require turkers to correctly identify a set of 48 dog breed classification questions. Independent two-sample t-tests, linear and logistic regression models, and randomization inference are used to perform inference on data collected from the running the first experimental design. Linear and logistic regression models, general linear hypothesis, weighted difference-in-means and simulation of potential outcome schedules are used to perform inference on data collected from the running the second experimental design. Results from both experiments fails to reject the null hypothesis that higher reward has zero effect on work quality nor worker performance. Through out the experiment, we observe that tasks attached with higher rewards tend to be claimed much sooner but no effect is observed on the returned task accuracies.

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 gives an overview of prior research conducted on the same topic, section 3 states our research hyopthesis and identification strategies, section 4 walks through the two

exerpimental designs and their motivations, detailing the platforms used and the experimental schedule, sections 5 and 6 present the data and analysis for the two experiment designs, section 6 discusses overall results, section 7 looks at the limitation of our project and suggest future lines of investigation, followed by a 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. A web link to our Github repo will be submitted along with the final paper.

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 turker performance or task quality. To operationalize this construct, we define the treatment variable as rewards 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 rewards are randomly assigned to each HIT. In each of the two experimental designs, a total of 48 image classifications questions on dog breeds are prompted in each HIT. The four levels of rewards 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 reward, controlling for other pre-treatment covariates for better precision. The resultant coefficient of assigned reward should be an estimate of the average treatment effect on turker performance or task quality. 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 48 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, 41 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 [1]. 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 sequence 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 [2]. Below is a print screen of the dog classication task page. For each question, eight different dog breed choices in additional to one "Not A Dog" choice as screener.

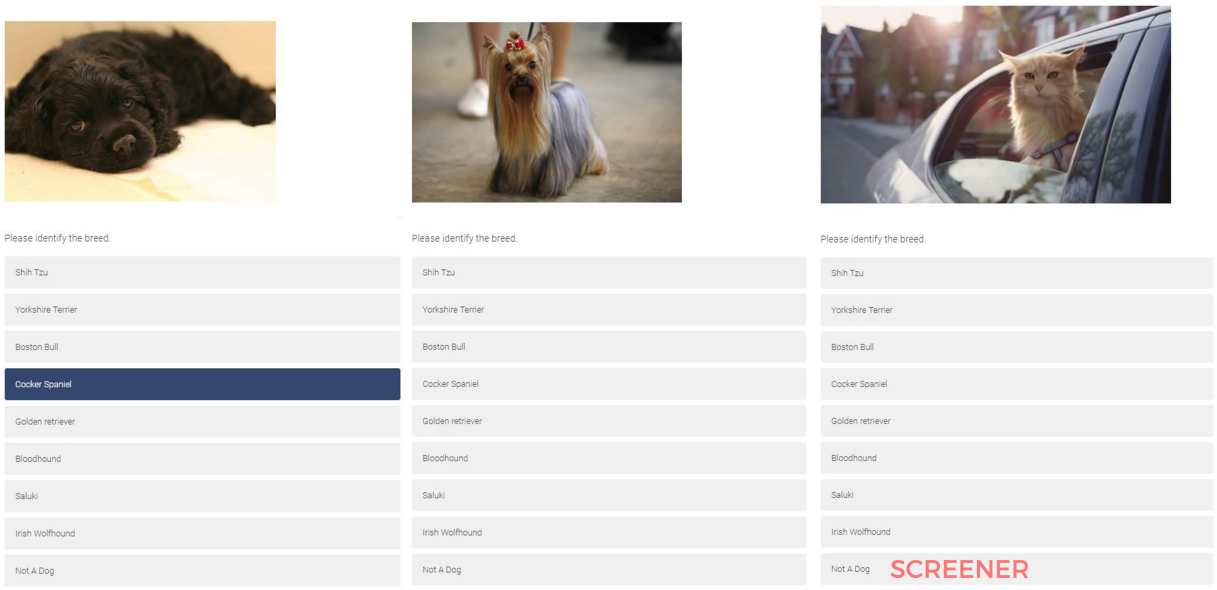


Figure 1:

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,

[1] Shih-Tzu: small and hairy. Yorkshire Terrier: small and hairy. Boston Bull: small and short hair. Cocker Spaniel: medium and hairy. Golden Retriever: medium and hairy. Bloodhound: big and short hair. Saluki: big and short hair. Iris Wolfhound: big and hairy.

[2] http://vision.stanford.edu/aditya86/ImageNetDogs/

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 publish order of postings to minimize publish 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. Note that in design 1, the treatment variable is defined as HIT pay rate displayed on the AMT platform, and the outcome variable is defined as returned HIT accuracy. Design 1 details are summarized below:

Experiment Schedule for Design 1

Publish Order	Date	Time Frame	Treatment (Pay Rate)	Available HITs
Pilot	Oct 28, 2017 (Saturday)	Morning	\$0.10	50
Pilot	Oct 29, 2017 (Sunday)	Afternoon	\$0.25	50
1	Nov 11, 2017 (Saturday)	Morning	\$0.10	100
1	Nov 11, 2017 (Saturday)	Afternoon	\$0.55	100
1	Nov 12, 2017 (Sunday)	Morning	\$0.25	100
1	Nov 12, 2017 (Sunday)	Afternoon	\$0.40	100
2	Nov 18, 2017 (Saturday)	Morning	\$0.40	100
2	Nov 18, 2017 (Saturday)	Afternoon	\$0.25	100
2	Nov 19, 2017 (Sunday)	Morning	\$0.55	100
2	Nov 19, 2017 (Sunday)	Afternoon	\$0.10	100

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 sequence 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. Note that in design 2, the treatment variable is defined as bonus rates that are assigned within the Qaultrics survey and not displayed on the AMT platform, and the outcome variable is defined as the turkers' performance. Design 2 details are summarized below:

Experiment Schedule for Design 2

Name	Date	Time Frame	Base pay rate	Treatments (bonuses)	Available HITs
Pilot	Nov 23, 2017 (Thursday)	All day	\$0.10	\$0.00, \$0.05, \$0.10, \$0.15	60
Main	Nov 26, 2017 (Sunday)	All day	\$0.22	\$0.00, \$0.10, \$0.20, \$0.30	240

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 design 2 results section will give describe attrition data in detail.

Finally, in both designs, we assume all subjects, whether an HIT or a turker, effectively receives treatment. We may not observe all turkers' outcomes in design 2, but that is an issue of attrition, not non-compliance. Given the mechanics of our experiment, an HIT can only be published if it is attached with a pay rate and a turker would only sign up an HIT knowing how much it will pay. We assume all the outcomes collected belongs to compliers. In design 2, we highlight with red and bolden the statement about bonuses at the top of each question session, if the turker is assigned with a session attached with a bonus, to make sure that turkers are aware of it. Therefore, we don't have specific concerns about non-compliance in this project.

5 Design 1 Results

5.1 Pilot Study (Design 1)

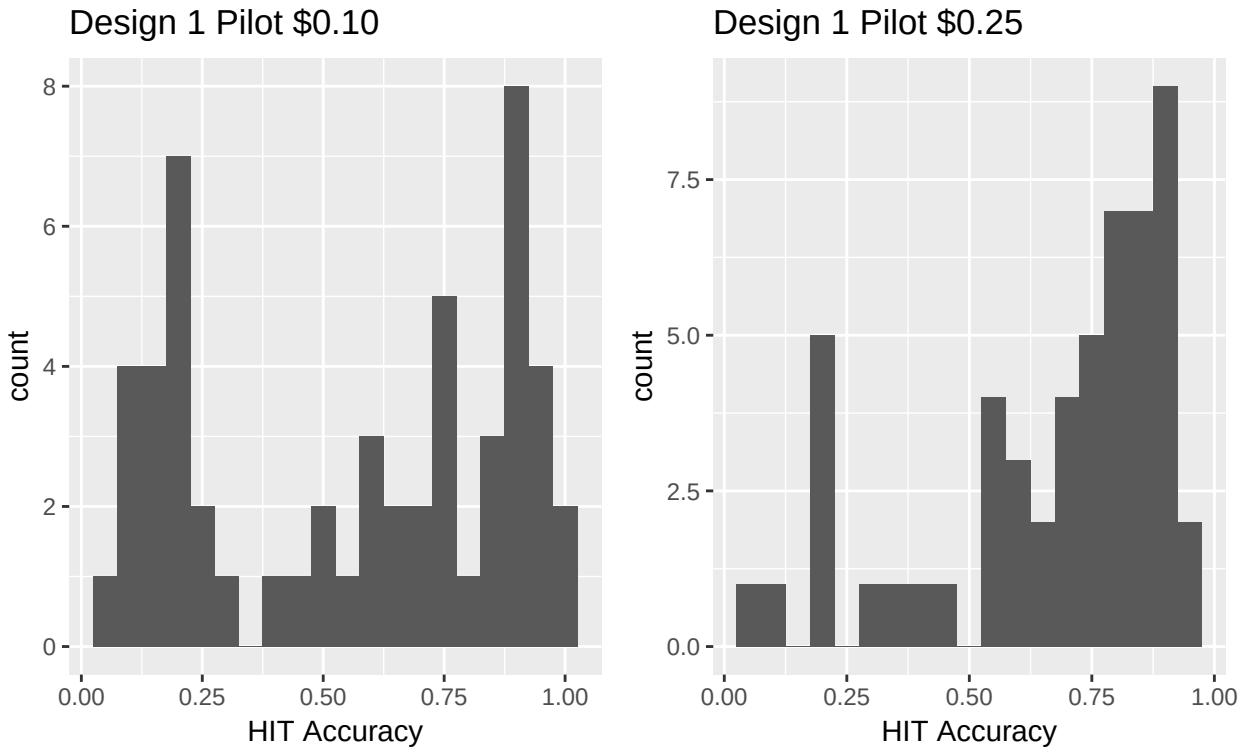
By running a pilot study for design 1, we tested the experiment protocol, identified problems in our AMT and Qualtrics workflow and conducted a power analysis on the collected data. During the last weekend of Oct 2017, we published two non-overlapping HIT postings each with 50 HITs available. One at 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
```

Robust Standard Errors are applied

```
## Simple Model with No Covariates:

## estimated average causal effect =  0.7555556
## robust standard error =  0.3659939
## 95% confidence interval =  0.02993712 1.481174
## p-value =  0.04142159

##
##
## Full Model with Covariates:

## estimated average causal effect =  0.6507679
## robust standard error =  0.3467159
## 95% confidence interval =  -0.03702313 1.338559
## p-value =  0.06341132
```

5.2 Main experiment (Design 1)

We ran the main Design 1 (D1) experiment over the second and third 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 publish orders of HIT postings (publish order1 is 0.10, 0.55, 0.25, 0.40; publish order 2 is 0.40, 0.25, 0.55, 0.10).

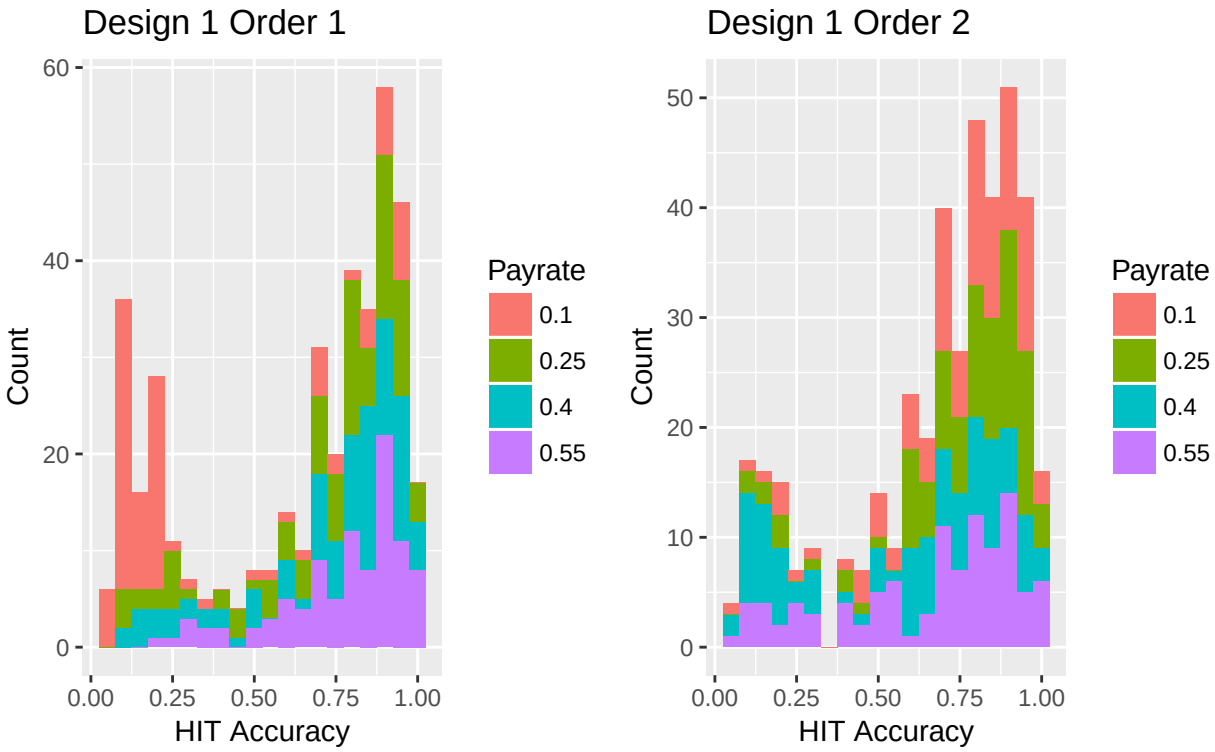
5.2.1 Data (Design 1 Main)

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. Publish 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 publish 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 publish 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 publish 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 publish 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 publish order 1, same results are only observed in the comparisons involving $0.10 HITs. With data from publish 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 publish 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
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 publish orders. Doing this gives us a p-value of 0.038, suggesting that the publish 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 seven 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 publish order and additionally detect any publish order 1 or first weekend effects. Model 4 further includes an interaction term between treatment and publish order to look for heterogeneous treatment effects – whether receiving treatment in publish order 1 (first weekend) is better or worse than receiving the same treatment in publish 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. Model 7 adopts the specification of model 3 but switch out the outcome variable as time spent on task. Model 7 adopts the specification of model 4 but switch out the outcome variable as time spent on task. Note that model 1, 2, 3, 4, 6, 7 and 8 assume that the relationship between outcome and treatment is linear (first order treatment term).

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

(7) $time\ spent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$

(8) $time\ spent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$

Figure 2 shows the corresponding regression table. 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 and 6 estimated no significant treatment effects. Model 3 is our most precise guess for the ATE using pooled data without further assumption of heterogeneous treatment effects, nor non-linearity. It estimates that for the pooled publish order 1 and publish order 2 data combined, for every additional dollar spent on an HIT, model 3 estimates an increment in accuracy of 0.286. But because of our wide .95 uncertainty range between -0.296 and 0.869, we cannot be sure if this increment is significantly different than zero. In fact, paying more may even have a= negative impact on accuracy.

```
##
##  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
```

Design 1 Main ATE Estimation

	Dependent variable:							
	accuracy OLS						time_spent OLS	
	simple (1)	+ covariates (2)	+ order (3)	+ order interaction (4)	+ 2nd order (5)	+ cov interaction (6)	+ covariates (7)	+ order interaction (8)
treatment	0.288	0.288	0.286	-0.257*	0.584	0.247	-7.728	14.574***
	(0.325)	(0.318)	(0.297)	(0.103)	(0.727)	(0.264)	(12.449)	(1.220)
CQ1a lot more than half		0.130**	0.129**	0.119**	0.121**	0.174	-2.603	-2.185
		(0.042)	(0.041)	(0.038)	(0.038)	(0.121)	(4.205)	(4.326)
CQ1around half		-0.009	-0.010	-0.011	-0.008	0.008	1.847	1.850
		(0.016)	(0.015)	(0.014)	(0.014)	(0.039)	(4.467)	(4.520)
CQ2_3		-0.018	-0.020	-0.024	-0.025	0.014	-2.687	-2.545
		(0.017)	(0.015)	(0.015)	(0.015)	(0.019)	(2.002)	(2.064)
CQ3No		0.084	0.087	0.088	0.094	0.179	3.343	3.285
		(0.063)	(0.064)	(0.062)	(0.064)	(0.102)	(5.337)	(5.631)
CQ3Yes		0.168***	0.174***	0.165**	0.167**	0.232***	0.315	0.672
		(0.050)	(0.050)	(0.051)	(0.051)	(0.062)	(4.519)	(4.650)
order1			-0.049	-0.406***	-0.405***	-0.400***	-0.509	14.125***
			(0.087)	(0.119)	(0.092)	(0.115)	(3.444)	(3.574)
I(treatment2)					-1.294			
					(1.092)			
treatment:order1				1.099***	1.097***	1.089***		-45.129***
				(0.270)	(0.238)	(0.259)		(10.818)
treatment:CQ1a lot more than half						-0.170		
						(0.280)		
treatment:CQ1around half						-0.054		
						(0.092)		
treatment:CQ2_3						-0.118		
						(0.063)		
treatment:CQ3No						-0.289		
						(0.356)		
treatment:CQ3Yes						-0.213		
						(0.255)		
Constant	0.561***	0.405*	0.430**	0.624***	0.523***	0.463***	57.765***	49.804***
	(0.142)	(0.173)	(0.141)	(0.066)	(0.130)	(0.090)	(7.362)	(7.646)
Observations	817	817	817	817	817	817	817	817
R2	0.028	0.118	0.126	0.226	0.236	0.231	0.005	0.011
Adjusted R2	0.027	0.112	0.118	0.218	0.228	0.219	-0.004	0.001
Residual Std. Error	0.285 (df = 815)	0.273 (df = 810)	0.272 (df = 809)	0.256 (df = 808)	0.254 (df = 807)	0.256 (df = 803)	48.147 (df = 809)	48.028 (df = 808)
F Statistic	23.286*** (df = 1; 815)	18.118*** (df = 6; 810)	16.594*** (df = 7; 809)	29.480*** (df = 8; 808)	27.703*** (df = 9; 807)	18.553*** (df = 13; 803)	0.533 (df = 7; 809)	1.095 (df = 8; 808)

Note: *p<0.05; **p<0.01; ***p<0.001

Figure 2: 16

```
##
##  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
```

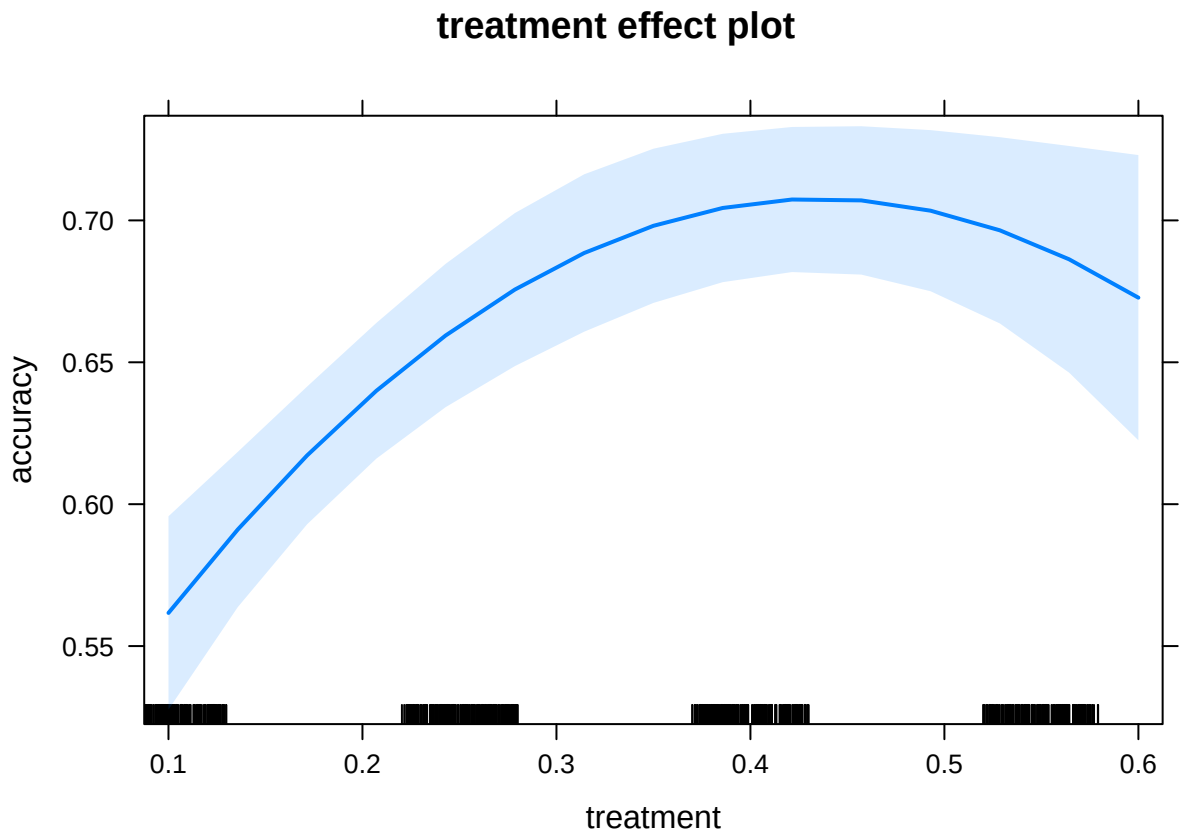
In model 4, estimated ATE for publish order 1 is significant with a point estimate of \$0.8423, with a .95 confidence interval between 0.348 and 1.3363. That is, for every additional \$0.10 spent on an HIT, we should expect an increase in accuracy between 0.0348 and 0.13363. Estimated ATE for publish 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, 4 ATEs for model 7-8), we apply Bonferroni correction to adjust the critical cut-off level to 0.00385. After this adjustment, the estimated ATE is not significant anymore. Notice that in model 4 the best estimate for ATE in publish order 1 and publish order 2 are off opposite signs, coefficient for publish order 1 is also highly significant, reinforcing our earlier observation of pulbish order or weekend effect.

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

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

In model 5, neither first order nor second order treatment regressors 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 publish order 1 is positive but that for publish order 2 is negative.

```
##
##  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
```



In model 6, estimated ATE for publish order 1 is significant even after Bonferroni correction is applied. It is estimated to be an 0.857 increase in accuracy for every additional dollar spent, that is 0.0858 for every addition $0.10 spent. This estimate can range from 0.316 to 1.397 with .95 confidence. The estimated ATE for publish order 2 is significant on its own but not after Bonferroni adjustment is applied. Note that from the regression table, the coefficient estimated for the "order1" and "treatment*order1" variables are largely significant, indicating publish order or weekend effect again.

```
##
##  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 6 inference for coefficient covariate Order1:

##      Estimate    Std. Error       t value      Pr(>|t|)
## -0.4003557194  0.1150266569 -3.4805472946  0.0005273829

##
## Model 6 inference for coefficient covariate Treatment*Order1:

##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.089429e+00 2.592405e-01 4.202386e+00 2.937564e-05
```

In model 7, as we swtich out the outcome variable from accuracy to time spent per HIT in seconds, neither treatment nor any other coefficients are significant. Our best guess suggests a decrease in time spent on an HIT by -7.72808 seconds. But we are not certain if it is significantly different from zero because our uncertainty ranges from a decrease by 32.164 seconds to an increase by 16.708 seconds.

```
##
##  Model 7: time_spent: ATE
## estimated average causal effect =  -7.72808
## Clustered standard errors = 12.44898
## .95 CI with clustered SE = [ -32.1642 16.70804 ]
## p-value =  0.5349192
```

In model 8, we detect any heterogeneous treatment effect on time spent for the different publish orders. Results suggest that ATE for publish order 1 is negative while that for publish order 2 is positive. For very additional dollar spent on an HIT in publish order 1 or the first weekend, on average turkers spent 30.55 seconds shorter on the HIT, equivalent to 3.055 seconds shorter per $0.10 spent. For every additional dollar spent in publish order 2 or the second weekend, on average turkers spent 14.574 seconds longer on the HIT, equivalent to 1.4574 seconds longer per $0.10 spent. These estimates even though they are statistically significant, are not pratically significant.

```
##
##  Model 8: time_spent: ATE for Order 1
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated average causal effect =  -30.55

## Clustered standard errors = 10.52

## .95 CI with clustered SE = [ -51.19975 -9.900247 ]

## p-value = 0.00377

##
##  Model 8: time_spent: ATE for Order 2
```

```
## estimated average causal effect =  14.57424
## Clustered standard errors = 1.219789
## .95 CI with clustered SE = [ 12.17991 16.96857 ]
## p-value =  2.044743e-30
```

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 the sharp null hypothesis would dramatically 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 publish 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)*

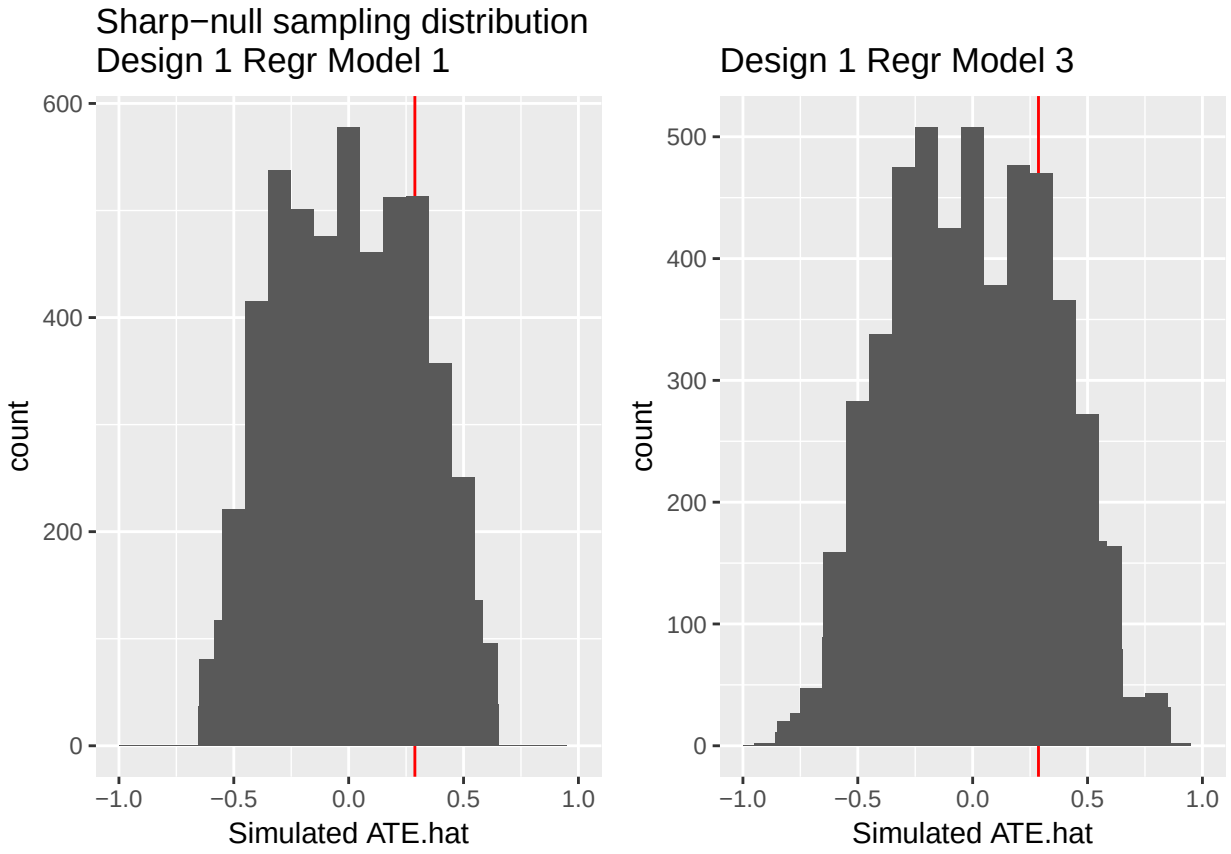
```
## Estimated ATE from simple regression (model 1): 0.288234
```

```
## p-value under sharp-null randomization inference: 0.4012
```

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

```
## p-value under sharp-null randomization inference: 0.4716
```

```
## `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 a similar inference procedure with our design 1 data, switching from linear regression to logistic regression for model 1 to 6, 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) $logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1 + \theta_8 * treatment^2 * cq1 + + \theta_9 * treatment^2 * cq1 + + \theta_{10} * treatment^2 * cq1$

Figure 3 shows the corresponding regression table. Clustered Standard Errors are applied

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 3.1 and model 4.1 only. Because the model 3.1 specification would more appropriately answer our question of ATE on pooled data (more general), without far-fetched assumption of non-linearity nor heterogeneous treatment effects that are based on our

Design 1 Main ATE Estimation

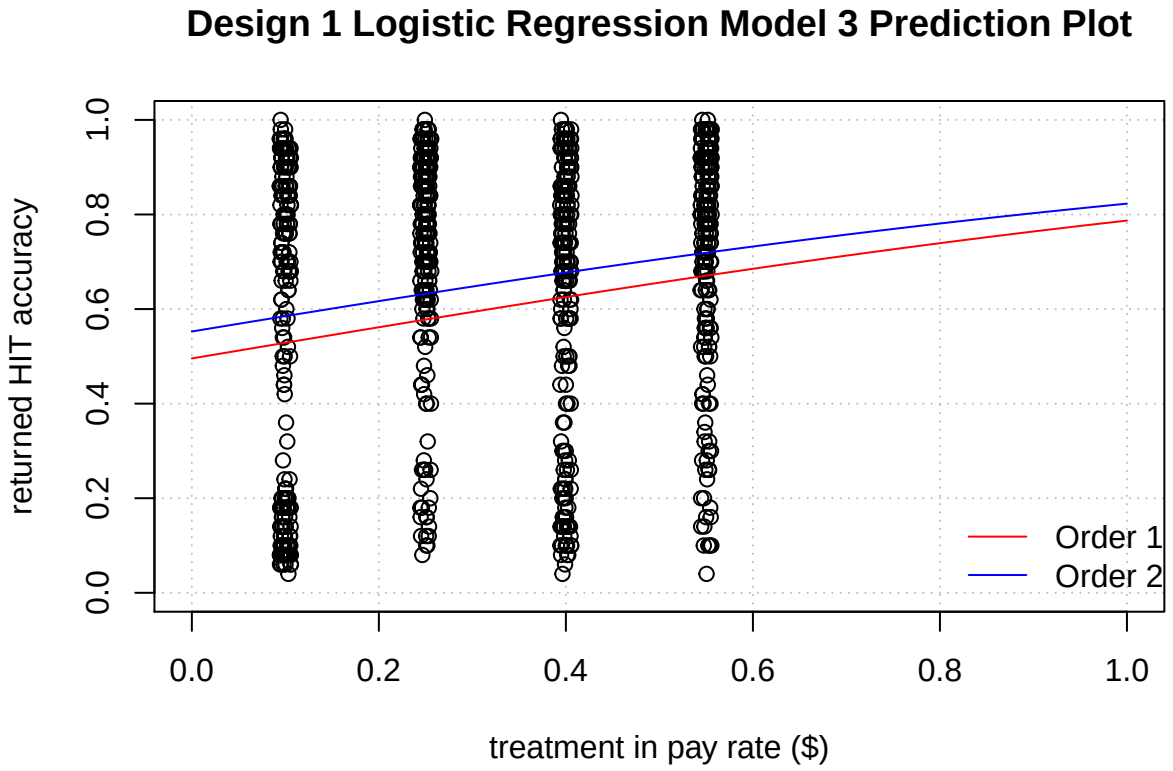
	Dependent variable:					
	accuracy logistic					
	simple (1)	+ covariates (2)	+ order (3)	+ interaction (4)	+ 2nd order (5)	+ cov interaction (6)
treatment	1.283	1.333	1.325	-1.226*	2.158	1.121
	(1.406)	(1.429)	(1.341)	(0.477)	(3.410)	(1.150)
CQ1a lot more than half		0.625***	0.619***	0.594***	0.607***	0.795
		(0.166)	(0.161)	(0.164)	(0.168)	(0.504)
CQ1around half		-0.041	-0.047	-0.048	-0.040	0.023
		(0.070)	(0.065)	(0.062)	(0.062)	(0.169)
CQ2_3		-0.083	-0.094	-0.110	-0.117	0.078
		(0.077)	(0.069)	(0.071)	(0.067)	(0.092)
CQ3No		0.345	0.358	0.383	0.406	0.824
		(0.270)	(0.274)	(0.276)	(0.287)	(0.493)
CQ3Yes		0.711**	0.742***	0.726**	0.735**	1.051***
		(0.220)	(0.221)	(0.232)	(0.237)	(0.275)
order1			-0.228	-1.862***	-1.832***	-1.853***
			(0.396)	(0.488)	(0.384)	(0.481)
I(treatment2)					-5.159	
					(5.108)	
treatment:order1				5.187***	5.061***	5.166***
				(1.167)	(1.013)	(1.137)
treatment:CQ1a lot more than half						-0.619
						(1.186)
treatment:CQ1around half						-0.209
						(0.408)
treatment:CQ2_3						-0.597
						(0.319)
treatment:CQ3No						-1.374
						(1.632)
treatment:CQ3Yes						-1.027
						(1.158)
Constant	0.230	-0.431	-0.316	0.565	0.152	-0.178
	(0.601)	(0.760)	(0.618)	(0.299)	(0.600)	(0.382)
Observations	817	817	817	817	817	817
Log Likelihood	-477.079	-456.748	-453.580	-423.436	-421.933	-422.050
Akaike Inf. Crit.	958.157	927.495	923.161	864.873	863.867	872.101
Note:						*p<0.05; **p<0.01; ***p<0.001

Figure 3:

prior research hypothesis, the model 4.1 specification let us investiagte the conspicuous abnormality observed in the differential outcome distribution in publish order 1 versus publish order 2, again without further assumption of non-linearity nor other heterogeneous treatment effects. Also, recall the interpretation with the linear models, the ATE estimated by linear model 3 was largely insignificant, while the ATE estimated by linear model 4 on publish order 1 data was the only significant effect after Bonferroni adjusted cut-off was applied.

Logistic regression model 3.1 estimates an insignificant coefficient of treatment as 1.325 with a wide confidence interval between -1.307 and 3.957. Taking the logit transformation of the outcome variable into account, we interpret the estimated odds of achieving full accuracy(1.00) to change by $e^{(1.325)(0.10)} = 1.1417$ times with every additional \$0.10 invested on an HIT, holding all other variables constant. The plot below shows how predicted accuracy varies with different pay rates, note that the predicted accuracy is not linear and bounded by 0 and 1.

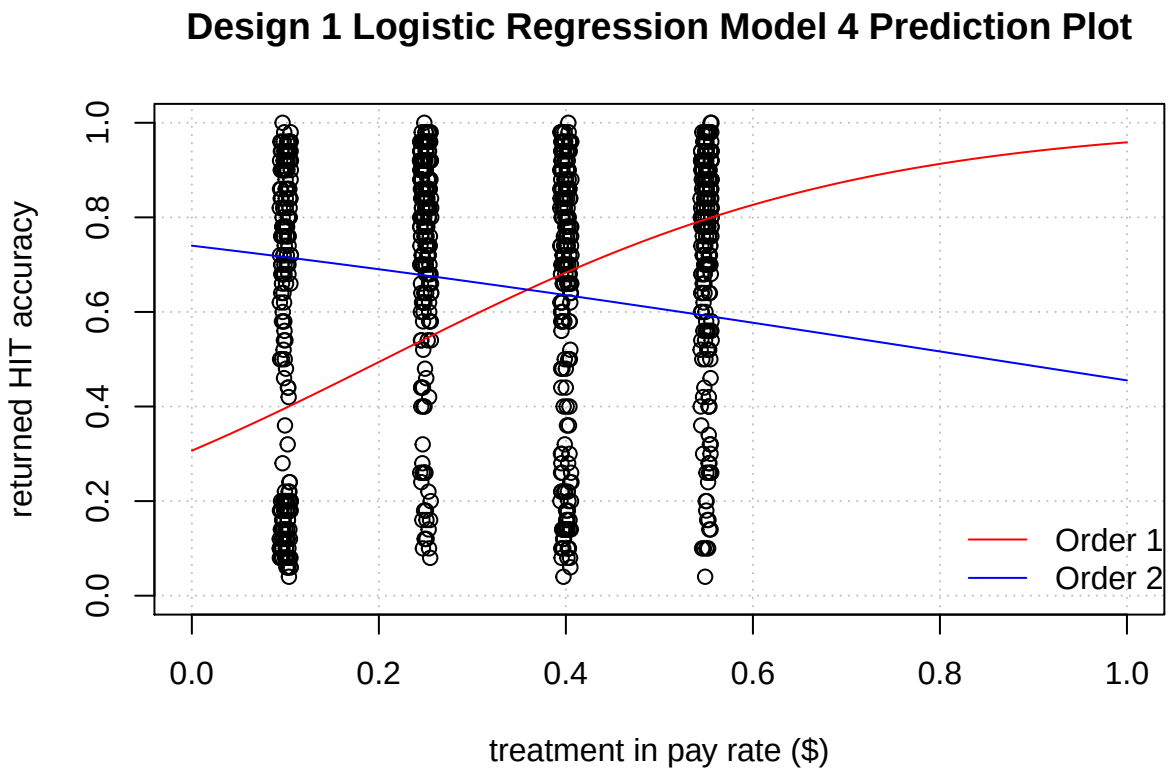
```
##
##  Model 3: Logistic Regression
## (derived using Simultaneous Tests for General Linear Hypotheses)

## estimated coefficient of treatment =  1.325

## Clustered standard errors = 1.341

## .95 Wald CI with clustered SE = [ -1.307 3.957 ]

## p-value = 0.323
```



Logistic regression model 4.1 estimates a significant linear combination of publish order 1 treatment as 3.961 with a confidence interval between -1.867 and 2.937. Taking the logit transformation of the outcome variable into account, we interpret the estimated odds of achieving full accuracy(1.00) to change by $e^{(1.325)(0.10)} = 1.486$

times with every additional $0.10 invested on an HIT, holding all other variables constant. Estimation of linear combination for publish order 2 treatment is again, like linear regression model 4, significant on its own but not any more when Bonferroni cut-off 0.00556 is applied. The plot below shows how predicted accuracy varies for both publish orders with different pay rates.

```
##
##  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
```



5.3 Design 1 Conclusion

To provide an overview for the design 1 results, the estimated effect of pay rate on HIT accuracy is insignificant when all data is pooled together. When publishing order is accounted for, we observe significant heterogeneous treatment effect and baseline difference. The same significance pattern is observed for both linear and logistic regression models. Randomization inference, by simulating the sampling distribution under the sharp-null, reinforces the linear regression result that the estimated ATE is insignificant with a more conservative p-value. Lastly, time spent per HIT is considered as a secondary outcome. The corresponding ATE is estimated to be negative but both statistically and practically insignificant. When publishing order is accounted for, we again observe statistically significant but practically insignificant heterogeneous treatment effect and baseline difference.

Therefore, with our best estimate, we fail to reject the null hypothesis that average treatment effect of pay rate on accuracy or time spent per HIT is zero. Rather, there are strong evidence of unaccounted effects which differentiated outcomes of the two publishing orders. For instance, the kind of turkers who sign up for low paying HITs on Saturday mornings may invest less effort on HITs than the kind who sign for the same HITs on Sunday afternoon, but our available time frame and budget prevented us from running experiments on enough weekends to determine such effect. Perhaps, the November 11th (Veteran's day) weekend is different from the November 18th weekend for American turkers, which is again unlikely because turkers came from multiple countries and we published our HITs on UTC. Finally, we investigated if publish order 1 has exhausted the pool of turkers or affected the baseline of turkers who participated in publish order 2. On its AMT platform, Amazon highlighted its service for providing access to over 500,000 turkers from over 190 countries. Even if all of them work part-time, and only 5% of them are active, 25,000 turkers are still accessible to us. It is unlikely that recruiting 749 turkers (405 who completed the HIT and 344 who viewed but left the HIT), that is less than 3% of 25,000 turkers or 0.1% of proclaimed 500,000 turkers, would have shifted the baseline for turkers recruited in publish order 2. Being wary of this result, we proceed to the analysis of design 2 in which subjects are defined as turkers rather than as HITs, and randomization occurs only after turkers signed up for an HIT. The experiment was also launched in a single posting on a single weekend. As such, selection bias, publish orders or weekend effects will not be of concern.

6 Design 2 Results

In design 2, all turkers begin by signing up for the same HIT base rate on the AMT platform with no knowledge that bonuses will be assigned to some of them. Then turkers are randomized into one of these four groups with equal probability. The randomization procedure only guarantees this probability for each new turker who sign up the HIT is the same, but not balanced assignment across all four groups. This process is similar to flipping a tetrahedral dice for each turker. To clarify, below are the definition of different group names, corresponding to the three sequentual sessions, each consist of 16 dog breed classification questions and one cat screener question.

- Group CCC : Turkers in this group do not receive bonuses in any sessions
- Group CCT : Turkers in this group only receive bonuses in the third session
- Group CTT : Turkers in this group only receive bonuses in the second and third sessions
- Group TTT : Turkers in this group receive bonsues in all three sessions

6.1 Pilot Study (Design 2)

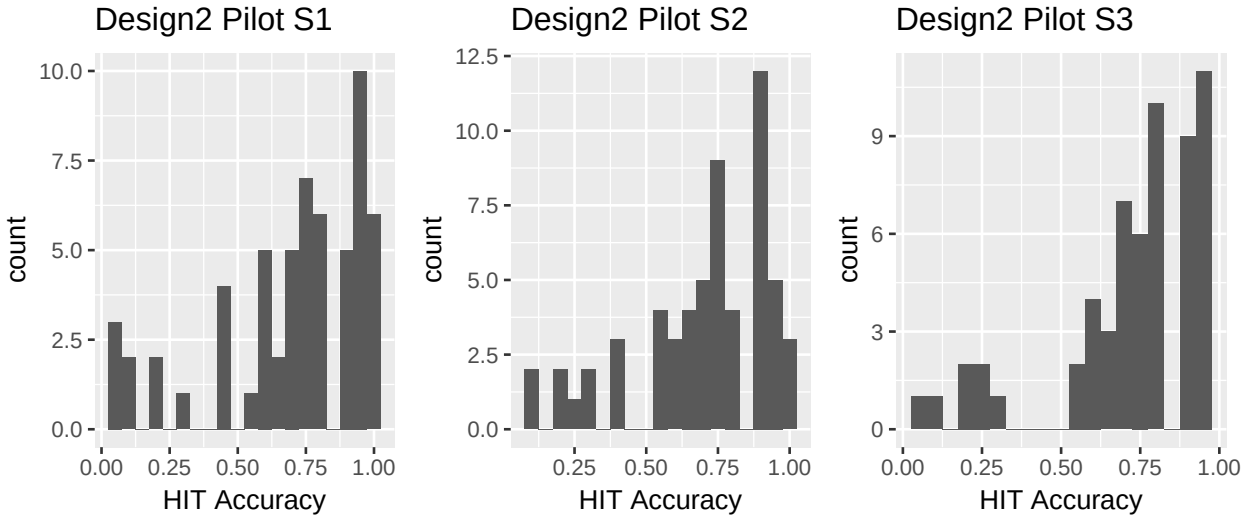
Pilot study for design 2 serve similar purpose for its main experiment in design 1, we identified technical problems by testing running our design and performed some preliminary power analysis.On Sunday, November 26 2017, we published one HIT posting of base rate $0.10 with 60 HITs available, each treatment session is attached with a bonus rate of $0.05. Our randomization procedure worked out using built-in functionalities in Qualtric as expected, giving us around 15 responses in each treatment group.

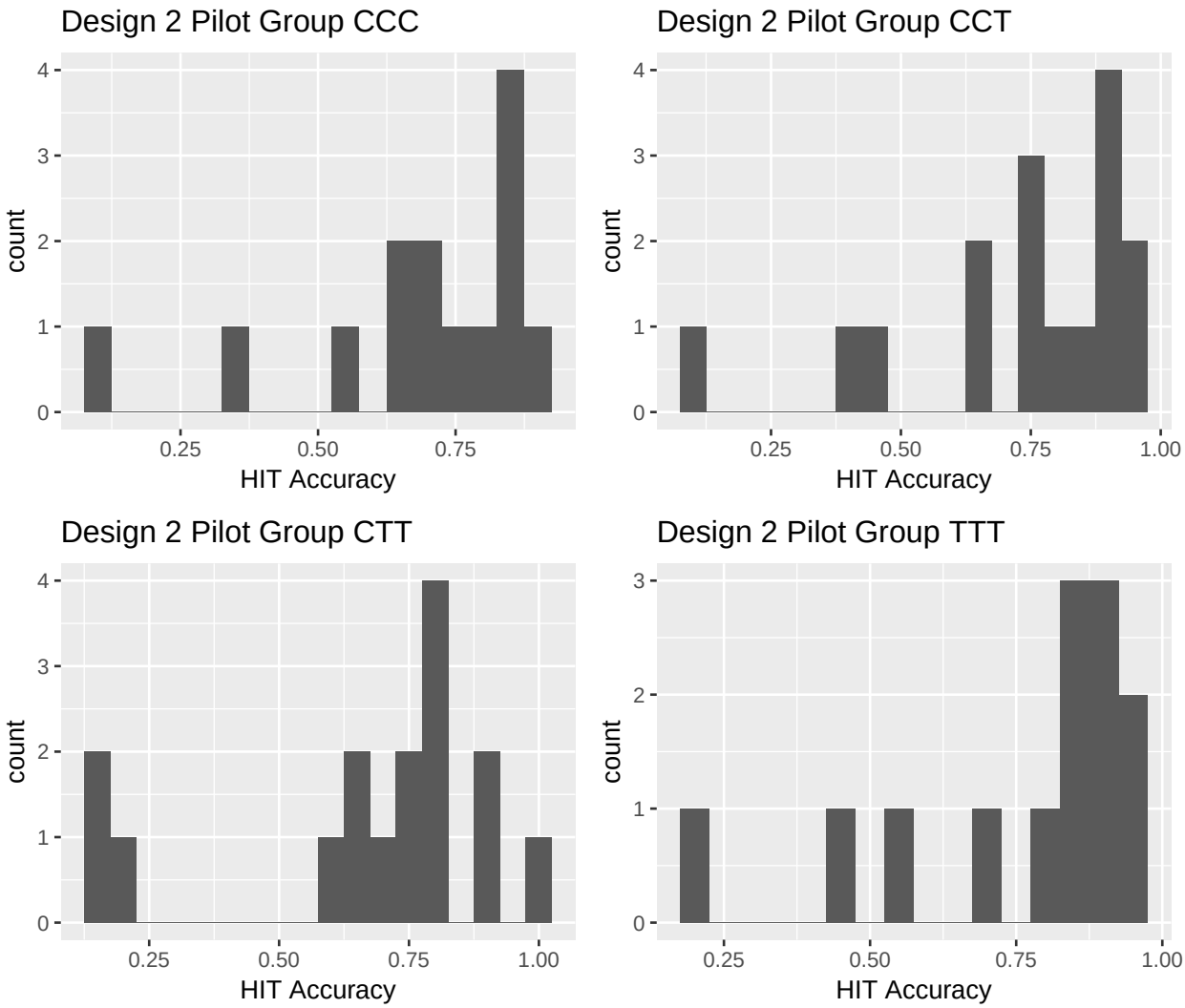
6.1.1 Data (Design 2 Pilot)

We observe 12 attriters, those refer to turkers who viewed by left the task in the middle. Their unobserved outcome is in roughly 1:5 ratio to the observed outcomes.

Group	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
All	/	59	20h 55min	6.651min	0.707	0.235
CCC	0.00	14	/	7.685min	0.681	0.217
CCT	$0.05	16	/	5.549min	0.729	0.236
CTT	$0.10	16	/	8.199min	0.663	0.27
TTT	$0.15	13	/	4.989min	0.763	0.219
Attritiers		12	/	/	/	/

Distribution of accuracy by session, or by treatment group are similarly right skewed. The bimodal distribution we observed in design 1 is much milder here.





6.1.2 Covariate Balance (Design 2 Pilot)

In terms of covariate balance, none of the treatment groups manage to predit the covariates, or in this case, the responses to any of our aptitude questions. Note that in design 2, because treatment assignment only happens after a turker has completed and submitted the aptitude questions, all five responses can be use as convariate controls for better precision.

```
##
## Covariate Balance Check Design 2 Pilot
## ================================================================================================
##                                                Dependent variable:
##                                 ------------------------------------------------------------
##                                  CQ1.1    CQ1.2     CQ1.3     CQ2      CQ3.1     CQ3.2    CQ3.3
##                                    (1)      (2)       (3)      (4)       (5)       (6)      (7)
## ------------------------------------------------------------------------------------------------
## groupCCT                        -0.089    0.063     0.027    0.259     0.223    -0.152   -0.071
##                                 (0.147)  (0.195)   (0.179)  (0.310)   (0.145)   (0.145)  (0.145)
##
## groupCTT                         0.098    0.063    -0.161   -0.366    -0.027    -0.089    0.116
```

```
##                                (0.171) (0.195)  (0.157) (0.260)  (0.179)  (0.179) (0.179)
##
## groupTTT                      -0.137   0.038    0.099   0.379    0.055   -0.214   0.159
##                               (0.143) (0.208)  (0.196) (0.319)  (0.181)  (0.181) (0.181)
##
## Constant                       0.214  0.500*** 0.286*  1.929*** 0.714***  0.214   0.071
##                               (0.118) (0.144)  (0.130) (0.202)  (0.130)  (0.130) (0.130)
##
## ------------------------------------------------------------------------------------------
## Observations                    59       59      59       59       59      59      59
## R2                             0.054   0.003    0.046   0.126    0.059    0.064   0.080
## Adjusted R2                    0.002   -0.052  -0.006   0.078    0.007    0.013   0.030
## Residual Std. Error (df = 55)  0.392   0.515    0.450   0.787    0.416    0.303   0.321
## F Statistic (df = 3; 55)       1.042   0.048    0.889   2.642    1.145    1.254   1.602
## ==========================================================================================
## Note:                                                        *p<0.05; **p<0.01; ***p<0.001
```

```
##
## Covariate Balance Check Design 2 Pilot
## ==========================================================================================
##                                                  Dependent variable:
##                              ------------------------------------------------------------
##                               CQ4.1   CQ4.2   CQ4.3   CQ4.4   CQ4.5   CQ5.1    CQ5.2   CQ5.3
##                                (1)     (2)     (3)     (4)     (5)     (6)      (7)     (8)
## ------------------------------------------------------------------------------------------
## groupCCT                      -0.152   0.063   0.036   0.107  -0.054   -0.009   0.009   0.000
##                               (0.135) (0.065) (0.165) (0.153) (0.192) (0.194)  (0.194) (0.000)
##
## groupCTT                      -0.152   0.188   0.098  -0.018  -0.116   -0.009   0.009   0.000
##                               (0.135) (0.104) (0.171) (0.134) (0.189) (0.194)  (0.194) (0.000)
##
## groupTTT                      -0.137   0.154   0.093  -0.143   0.033   -0.110   0.033   0.077
##                               (0.143) (0.108) (0.182) (0.101) (0.207) (0.207)  (0.207) (0.080)
##
## Constant                       0.214  -0.000   0.214   0.143  0.429** 0.571*** 0.429** -0.000
##                               (0.118)         (0.118) (0.101) (0.142) (0.142)  (0.142) (0.000)
##
## ------------------------------------------------------------------------------------------
## Observations                    59      59      59      59      59       59      59      59
## R2                             0.044   0.060   0.008   0.065   0.013   0.007    0.001   0.061
## Adjusted R2                   -0.009   0.009  -0.046   0.014  -0.040   -0.047  -0.054   0.010
## Residual Std. Error (df = 55)  0.306   0.304   0.459   0.343   0.502   0.514    0.514   0.130
## F Statistic (df = 3; 55)       0.834   1.167   0.154   1.279   0.248   0.138    0.010   1.191
## ==========================================================================================
## Note:                                                        *p<0.05; **p<0.01; ***p<0.001
```

CQ1 to CQ3 has the same interpretation as in design 1, CQ4_1 indicates if the turker is in age group 0 to 10, CQ4_2 indicates that of 11 to 20, CQ4_3 indicates that for 21 to 30, CQ4_4 indicates that for 31 to 40, CQ4_5 indicates that for 41 or more, CQ5_1 indicates the turker has a Linkedin account, CQ5_2 indicates the turker has no Linkedin account, CQ5_3 indicates if the turker has never heard of Linkedin

6.1.3 Treatment Effect Estimation (Design 2 Pilot)

For the ATE estimation in this pilot, we keep our analysis simple. That is, we only compare overall accuracies, total time spent per task and screeners accuracies across the four treatment groups. Granularity between the three different sessions is not investigated. The model specifications are listed here.

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

(full) $overall\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

(simple) $total\ timespent = \theta_0 + \theta_1 * treatment$

(full) $total\ timespent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

(simple) $all\ screeners\ passed = \theta_0 + \theta_1 * treatment$

(full) $all\ screeners\ passed = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

The regression table in Figure 4 shows that most estimated coefficients are not significant (Robust Standard Errors are applied). For many treatment group related coefficients, the size of standard error is very similar to that of the point estimate itself. Judging by the overall results of design 2 pilot, we decided to raise our HIT base rate from \$0.10 to \$0.22, and our bonus rate from \$0.05 to \$0.10 for the main experiment.

```
##
## Design 2 Pilot, Simple Model : Effect of being in group CCT rather than CCC on accuracy

## estimated average causal effect =  0.0484944
## robust standard error =  0.08287031
## 95% confidence interval =  -0.1175814 0.2145702
## p-value =  0.5608172

##
## Design 2 Pilot, Simple Model : Effect of being in group CTT rather than CCC on accuracy

## estimated average causal effect =  -0.01768207
## robust standard error =  0.08916341
## 95% confidence interval =  -0.1963695 0.1610054
## p-value =  0.8435331

##
## Design 2 Pilot, Simple Model : Effect of being in group TTT rather than CCC on accuracy

## estimated average causal effect =  0.08252532
## robust standard error =  0.08371201
## 95% confidence interval =  -0.0852373 0.2502879
## p-value =  0.3285365

##
## Design 2 Pilot, Full Model : Effect of being in group CCT rather than CCC on accuracy

## estimated average causal effect =  0.07670949
## robust standard error =  0.07192245
## 95% confidence interval =  -0.06824067 0.2216597
## p-value =  0.2919887

##
## Design 2 Pilot, Full Model : Effect of being in group CTT rather than CCC on accuracy

## estimated average causal effect =  -0.03165037
## robust standard error =  0.09657354
## 95% confidence interval =  -0.2262816 0.1629808
## p-value =  0.7446677
```

Design 2 Pilot ATE Estimation

	Dependent variable:					
	overall_accuracy		total_timespent		all_screeners_passed	
	OLS		OLS		OLS	
	simple	full	simple	full	simple	full
	(1)	(2)	(3)	(4)	(5)	(6)
groupCCT	0.048	0.077	-128.142	-129.006	0.009	0.001
	(0.083)	(0.072)	(79.057)	(84.005)	(0.095)	(0.080)
groupCTT	-0.018	-0.032	30.822	12.477	0.009	-0.009
	(0.089)	(0.097)	(109.850)	(121.958)	(0.095)	(0.092)
groupTTT	0.083	0.072	-161.744*	-159.677*	-0.005	0.0004
	(0.084)	(0.084)	(79.485)	(81.387)	(0.105)	(0.084)
CQ1around half		0.020		-20.192		-0.160
		(0.081)		(81.411)		(0.086)
CQ1a lot more than half		0.147		121.500		-0.010
		(0.081)		(110.526)		(0.059)
CQ2		-0.082		-60.693		-0.058
		(0.043)		(48.974)		(0.051)
CQ3Maybe		0.023		9.883		0.067
		(0.129)		(138.300)		(0.164)
CQ3Yes		-0.080		25.893		0.170
		(0.111)		(105.004)		(0.137)
CQ411 to 20		0.079		73.590		0.094
		(0.157)		(102.600)		(0.117)
CQ421 to 30		0.015		74.297		0.020
		(0.145)		(102.873)		(0.153)
CQ431 to 40		0.068		40.291		0.003
		(0.117)		(90.443)		(0.102)
CQ441 or more		0.084		9.954		0.033
		(0.119)		(83.288)		(0.107)
CQ5No		-0.230		-84.898		-0.098
		(0.118)		(90.491)		(0.112)
CQ5Yes		-0.331**		-3.310		-0.131
		(0.111)		(96.042)		(0.137)
Constant	0.681***	1.078***	461.098***	545.253**	0.929***	1.085***
	(0.058)	(0.194)	(70.947)	(190.146)	(0.071)	(0.183)
Observations	59	59	59	59	59	59
R2	0.028	0.282	0.112	0.221	0.001	0.168
Adjusted R2	-0.025	0.054	0.064	-0.027	-0.054	-0.097
Residual Std. Error	0.238 (df = 55)	0.229 (df = 44)	237.059 (df = 55)	248.299 (df = 44)	0.260 (df = 55)	0.266 (df = 44)
F Statistic	0.527 (df = 3; 55)	1.237 (df = 14; 44)	2.319 (df = 3; 55)	0.891 (df = 14; 44)	0.011 (df = 3; 55)	0.634 (df = 14; 44)
Note:						*p<0.05; **p<0.01; ***p<0.001

Figure 4:

```
##
## Design 2 Pilot, Full Model : Effect of being in group TTT rather than CCC on accuracy
## estimated average causal effect =  0.07211621
## robust standard error =  0.0838721
## 95% confidence interval =  -0.0969169 0.2411493
## p-value =  0.3945405
```

6.2 Main experiment (Design 2)

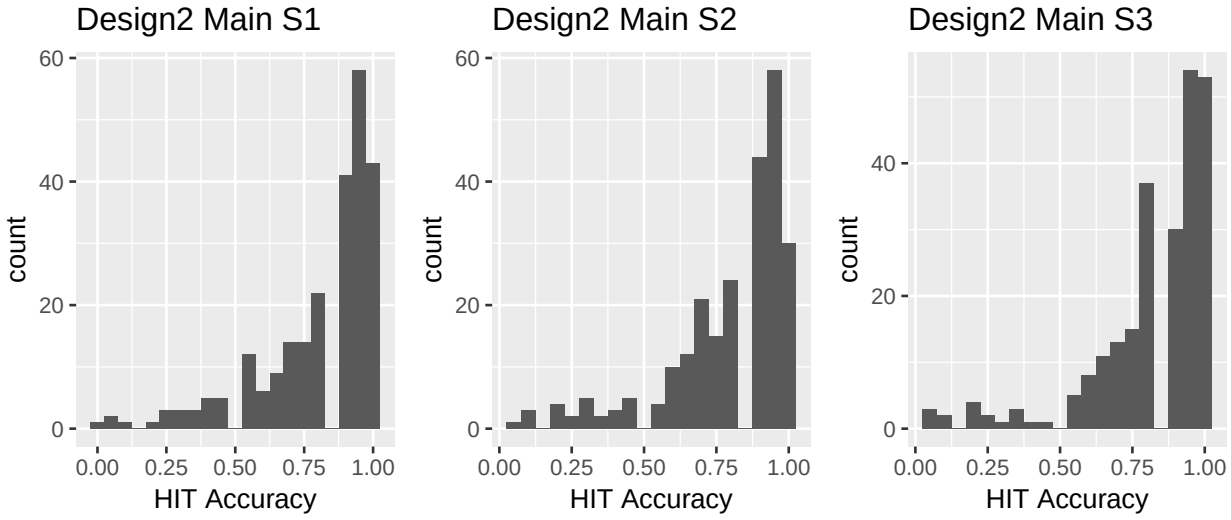
We ran the main Design 2 (D2) experiment on Sunday November 26 2017. We publised one posting of base rate $0.22 with 240 HITs available, each session is attached with a bonus rate of $0.10.

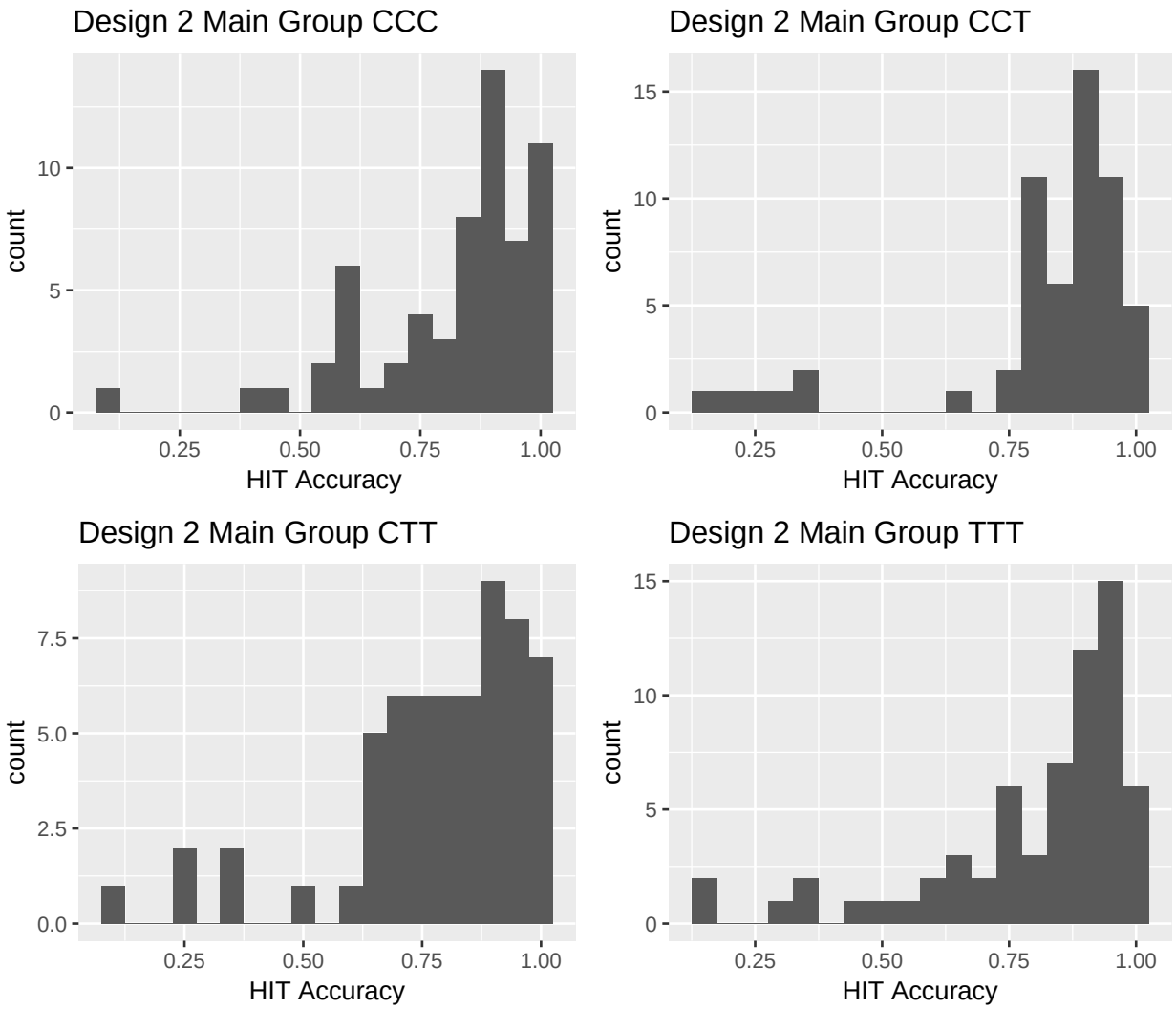
6.2.1 Data (Design 2 Main)

The main experiment, because of the higher base rate, took less than half of the time in the pilot for all HITs to be claimed. We observe 49 attriters, which correspond to roughly 1:5 observed outcomes to unobserved outcomes again. At a first glance, the accuracy of each treatment group don't differ noticeably.

Group	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
All	/	243	10h 2min	6.239min	0.807	0.197
CCC	0.00	61	/	5.854min	0.823	0.179
CCT	$0.05	58	/	6.22min	0.821	0.205
CTT	$0.10	60	/	5.834min	0.783	0.199
TTT	$0.15	64	/	7.004min	0.8	0.206
Attritiers		49	/	/	/	/

The distribution of accuracy by group or by session is again right skewed and very similar to that observed in the pilot.





6.2.2 Differential Attrition (Design 2 Main)

During the execution of design 1, we saw that more turkers who signed up for lower pay HITs tend to view the HITs but not complete them. While it was not a problem of design 1 because our unit of analysis was defined as a returned HIT, Design 2 is very sensitive to attrition. Here the unit of analysis is defined as a single turker. Therefore, turkers who viewed but left the HITs in the middle are counted as attriters. We also saw a roughly 1:5 attriters(outcome not observed) to completers(outcome observed) ratio in both the design 2 pilot and main experiment data. Below, we regress whether outcome is observable on treatment group assignment and in additional look for differential attrition using all the covariates we have. The regression results suggests that treatment group assignment is not a significant predictor of attrition. But if we look among the covariates, 4 out of 5 of the significant coefficients belong to terms or interaction terms that involves "CQ1a lot more than half".

Specifically in model 2, when treatment groups are not included, turkers are 13.3% less likely to attrit if a lot more than half of their friends own dogs compared to having less of half of their friends own dogs. They are 20.1% less likely to attrit if they have lived with a dog before or seriously considered owning a dog, compared to answering no to either. In layman's term, more exposures to or life experiences with dog may be positively associated with how comfortable the turkers feel about the task. On the other hand, model 3 estimates that turkers assigned in group CCC are 44.6% less likely to attrit if a lot more than half of their friends own dogs

and 30.5% less likely to attrit if around half of their friends own dogs, compared to having a lot less than half of their friends own dogs. Model 3 also estimates that for those who have half of their friends being dog owners, being assigned to group CCT makes them 64.3% more likely to attrit and being assigned to group TTT makes them 39.6% more likely to attrit. In layman's term, more regular exposures to dogs is associated with a turker being less likely to attrit if no bonuses are assigned. But the opposite happens when bonuses are assigned. This differential attrition interpretation is a little peculiar to us. Nevertheless, differential attrition remains at threat to unbiasedness of design 2 and we will discuss improvements for future line of investigation in section 7.

```
##
## Differential Attrition Check Design 2 Main
## ================================================================================================
##                                                Dependent variable:
##                                    ------------------------------------------------------------
##                                                       attrited
##                                          (1)                (2)                (3)
## ------------------------------------------------------------------------------------------------
## groupCCT                                0.031                                 -0.552
##                                        (0.060)                               (0.592)
##
## groupCTT                                0.014                                 -0.504
##                                        (0.059)                               (0.340)
##
## groupTTT                               -0.030                                 -0.018
##                                        (0.059)                               (0.359)
##
## CQ1around half                                            -0.070             -0.305*
##                                                           (0.063)            (0.133)
##
## CQ1a lot more than half                                   -0.133*            -0.446***
##                                                           (0.064)            (0.130)
##
## CQ2                                                        0.016              0.052
##                                                           (0.029)            (0.056)
##
## CQ3Maybe                                                  -0.178              0.110
##                                                           (0.112)            (0.249)
##
## CQ3Yes                                                    -0.201**           -0.037
##                                                           (0.076)            (0.149)
##
## CQ411 to 20                                                0.109             -0.087
##                                                           (0.106)            (0.193)
##
## CQ421 to 30                                                0.125              0.121
##                                                           (0.100)            (0.172)
##
## CQ431 to 40                                                0.180              0.159
##                                                           (0.107)            (0.183)
##
## CQ441 or more                                              0.071             -0.052
##                                                           (0.093)            (0.151)
##
## CQ5No                                                      0.084              0.062
```

```
##                                                    (0.254)          (0.381)
##
## CQ5Yes                                              0.103            0.171
##                                                    (0.253)          (0.372)
##
## groupCCT:CQ1around half                                              0.405
##                                                                     (0.215)
##
## groupCTT:CQ1around half                                              0.257
##                                                                     (0.186)
##
## groupTTT:CQ1around half                                              0.282
##                                                                     (0.176)
##
## groupCCT:CQ1a lot more than half                                    0.643**
##                                                                     (0.208)
##
## groupCTT:CQ1a lot more than half                                     0.204
##                                                                     (0.179)
##
## groupTTT:CQ1a lot more than half                                    0.396*
##                                                                     (0.178)
##
## groupCCT:CQ2                                                        -0.033
##                                                                     (0.081)
##
## groupCTT:CQ2                                                        -0.054
##                                                                     (0.083)
##
## groupTTT:CQ2                                                        -0.082
##                                                                     (0.083)
##
## groupCCT:CQ3Maybe                                                   -0.639
##                                                                     (0.337)
##
## groupCTT:CQ3Maybe                                                    0.244
##                                                                     (0.386)
##
## groupTTT:CQ3Maybe                                                   -0.451
##                                                                     (0.356)
##
## groupCCT:CQ3Yes                                                     -0.393
##                                                                     (0.248)
##
## groupCTT:CQ3Yes                                                      0.089
##                                                                     (0.225)
##
## groupTTT:CQ3Yes                                                     -0.297
##                                                                     (0.197)
##
## groupCCT:CQ411 to 20                                                 0.395
##                                                                     (0.362)
##
## groupCTT:CQ411 to 20                                                 0.428
```

```
##                                                                                                  (0.281)
##
## groupTTT:CQ411 to 20                                                                              0.059
##                                                                                                  (0.310)
##
## groupCCT:CQ421 to 30                                                                              0.043
##                                                                                                  (0.358)
##
## groupCTT:CQ421 to 30                                                                             -0.078
##                                                                                                  (0.259)
##
## groupTTT:CQ421 to 30                                                                              0.143
##                                                                                                  (0.277)
##
## groupCCT:CQ431 to 40                                                                              0.316
##                                                                                                  (0.380)
##
## groupCTT:CQ431 to 40                                                                              0.154
##                                                                                                  (0.294)
##
## groupTTT:CQ431 to 40                                                                             -0.275
##                                                                                                  (0.299)
##
## groupCCT:CQ441 or more                                                                            0.279
##                                                                                                  (0.338)
##
## groupCTT:CQ441 or more                                                                            0.242
##                                                                                                  (0.237)
##
## groupTTT:CQ441 or more                                                                            0.084
##                                                                                                  (0.253)
##
## groupCCT:CQ5No                                                                                    0.345
##                                                                                                  (0.531)
##
## groupCTT:CQ5No                                                                                    0.230
##                                                                                                  (0.124)
##
## groupTTT:CQ5No                                                                                    0.083
##                                                                                                  (0.124)
##
## groupCCT:CQ5Yes                                                                                   0.319
##                                                                                                  (0.527)
##
## groupCTT:CQ5Yes
##
##
## groupTTT:CQ5Yes
##
##
## Constant                                          0.141***                  0.197                 0.291
##                                                    (0.042)                 (0.265)               (0.394)
##
## ------------------------------------------------------------------------------------------------------
```

```
## Observations                          284                 284                 284
## R2                                  0.004               0.072               0.235
## Adjusted R2                        -0.007               0.035               0.090
## Residual Std. Error          0.353 (df = 280)    0.346 (df = 272)    0.336 (df = 238)
## F Statistic                0.373 (df = 3; 280) 1.923* (df = 11; 272) 1.625* (df = 45; 238)
## ============================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

6.2.3 Covariate Balance (Design 2 Main)

In terms of covariate balance, none of the treatment groups manage to predit the covariates, or in this case, the responses to any of our aptitude questions. The results are similar to that observed in the pilot.

```
##
## Covariate Balance Check Design 2 Main
## ================================================================================
##                                        Dependent variable:
##                                -------------------------------------------------
##                                 CQ1.1   CQ1.2    CQ1.3     CQ2      CQ3.1    CQ3.2   CQ3.3
##                                  (1)     (2)      (3)      (4)      (5)      (6)     (7)
## --------------------------------------------------------------------------------
## groupCCT                       -0.027   0.006    0.021    0.068    -0.031   0.072  -0.041
##                                (0.054) (0.084)  (0.085)  (0.124)  (0.062)  (0.062) (0.062)
##
## groupCTT                        0.070   0.042    -0.113   -0.099   0.042   -0.014  -0.028
##                                (0.062) (0.084)  (0.083)  (0.127)  (0.055)  (0.055) (0.055)
##
## groupTTT                        0.068   0.078    -0.145   -0.080   -0.040   0.013   0.026
##                                (0.062) (0.084)  (0.082)  (0.123)  (0.062)  (0.062) (0.062)
##
## Constant                       0.127** 0.408*** 0.465*** 1.789*** 0.859***  0.042  0.099*
##                                (0.040) (0.059)  (0.060)  (0.088)  (0.042)  (0.042) (0.042)
##
## --------------------------------------------------------------------------------
## Observations                    284     284      284      284      284      284     284
## R2                              0.014   0.004    0.021    0.008    0.008    0.019   0.009
## Adjusted R2                     0.003   -0.007   0.010    -0.002   -0.002   0.008  -0.002
## Residual Std. Error (df = 280)  0.362   0.499    0.489    0.738    0.356    0.237   0.284
## F Statistic (df = 3; 280)       1.290   0.372    2.000    0.770    0.767    1.798   0.806
## ================================================================================
## Note:                                                  *p<0.05; **p<0.01; ***p<0.001
```

```
##
## Covariate Balance Check Design 2 Main
## ================================================================================================
##                                              Dependent variable:
##                                -----------------------------------------------------------------
##                                 CQ4.1   CQ4.2   CQ4.3    CQ4.4   CQ4.5    CQ5.1    CQ5.2    CQ5.3
##                                  (1)     (2)     (3)      (4)     (5)      (6)      (7)      (8)
## ------------------------------------------------------------------------------------------------
## groupCCT                       -0.070   0.044   0.060    0.002   -0.036   -0.151   0.150   0.0002
##                                (0.041) (0.058) (0.068)  (0.057) (0.085)  (0.082)  (0.083)  (0.020)
##
## groupCTT                       -0.042   0.028   0.056   -0.028   -0.014   -0.056   0.070   -0.014
```

```
##                                 (0.045) (0.057) (0.068)  (0.054) (0.085)  (0.084)  (0.084)  (0.014)
##
## groupTTT                       -0.057   0.026   0.025   -0.030   0.035    -0.034   0.048   -0.014
##                                 (0.043) (0.056) (0.065)  (0.053) (0.085)  (0.084)  (0.084)  (0.014)
##
## Constant                       0.099** 0.113** 0.169*** 0.127** 0.493*** 0.465*** 0.521***  0.014
##                                 (0.036) (0.038) (0.045)  (0.040) (0.060)  (0.060)  (0.060)  (0.014)
##
## ------------------------------------------------------------------------------------------------
## Observations                     284     284     284      284     284      284      284      284
## R2                              0.013   0.002   0.004    0.002   0.003    0.013    0.012    0.007
## Adjusted R2                     0.002  -0.009  -0.007   -0.008  -0.008    0.002    0.001   -0.003
## Residual Std. Error (df = 280)  0.231   0.346   0.405    0.318   0.503    0.491    0.493    0.084
## F Statistic (df = 3; 280)       1.229   0.199   0.341    0.208   0.247    1.213    1.140    0.676
## ================================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

Definition of covariate variables remains unchanged in all experiments. Please reference the design 2 pilot version.

6.2.4 Treatment Effect Estimation (Design 2 Main)

In this subsection, we use three different approaches to estimate the effect of receiving bonuses on a turker performance.

6.2.4.1 Between Groups Comparison (Design 2 Main)

In the first approach, we pool data from all the sessions together and simply observe the effect of belonging to different groups. The following regression models are constructed with the specifications listed below. Model 1 looks at the effect of being assigned to a different treatment group on the overall HIT accuracy. Model 2 further controls for the responses to the aptitude questions as covariates for better precision. Model 3 and 4 are similar to model 1 and 2, but applied logistic instead of linear regression to better approximate the distribution of accuracy values. Model 5 and 6 are similar to model 1 and 2, but switched out the outcome variable as time spent instead. We separately estimated two models using screener(cat questions) results as outcome, the result was again insignicant therefore not included here.

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

(2) $overall\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

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

(4) $logit(overall\ accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

(5) $total\ time\ spent = \theta_0 + \theta_1 * treatment$

(6) $total\ time\ spent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

Figure 5 shows the corresponding regression table. Robust Standard Errors are applied.

Looing at results from linear models 1 to 4, being assigned to any of the treatment groups rather than group CCC does not make any statistical differences in the overall accuracy. For the linear models 1 and 2, the magnitude of estimated effects are in the order of less than five percentage points, suggesting no practical significance either. In fact, most estimated effects are mildly negative. All the corresponding robust confidence levels cover zero.

```
##
## Model 1: simple: ATE being in group CCT rather than group CCC:
```

Design 2 Main experiment ATE Estimation

	Dependent variable:					
	overall_accuracy				total_timespent	
	OLS		logistic		OLS	
	simple	full	simple	full	simple	full
	(1)	(2)	(3)	(4)	(5)	(6)
groupCCT	-0.001	0.013	-0.010	0.084	21.946	24.713
	(0.036)	(0.033)	(0.243)	(0.242)	(31.790)	(34.618)
groupCTT	-0.039	-0.033	-0.249	-0.228	-1.232	-5.639
	(0.035)	(0.032)	(0.220)	(0.208)	(28.794)	(28.817)
groupTTT	-0.023	-0.002	-0.150	0.001	68.965	68.494
	(0.035)	(0.030)	(0.226)	(0.210)	(37.967)	(38.113)
CQ1around half		0.081		0.470		-48.134
		(0.047)		(0.255)		(33.606)
CQ1a lot more than half		0.115*		0.718**		-30.540
		(0.047)		(0.263)		(31.103)
CQ2		0.024		0.168		15.220
		(0.016)		(0.115)		(15.956)
CQ3Maybe		-0.153		-0.708		-141.388
		(0.087)		(0.406)		(82.260)
CQ3Yes		0.101*		0.602*		-100.141
		(0.050)		(0.267)		(53.943)
CQ411 to 20		-0.065		-0.445		-14.201
		(0.055)		(0.369)		(60.593)
CQ421 to 30		-0.030		-0.239		0.953
		(0.045)		(0.336)		(54.521)
CQ431 to 40		0.008		0.039		19.894
		(0.049)		(0.360)		(60.118)
CQ441 or more		-0.032		-0.265		-51.929
		(0.041)		(0.298)		(47.400)
CQ5No		0.181*		1.179**		161.226***
		(0.080)		(0.409)		(42.751)
CQ5Yes		0.114		0.719		156.301***
		(0.082)		(0.414)		(43.122)
Constant	0.823***	0.482***	1.534***	-0.495	351.246***	323.026***
	(0.023)	(0.097)	(0.158)	(0.476)	(22.820)	(64.328)
Observations	243	243	243	243	243	243
R2	0.007	0.213			0.024	0.079
Adjusted R2	-0.006	0.165			0.012	0.022
Log Likelihood			-83.451	-70.854		
Akaike Inf. Crit.			174.903	171.707		
Residual Std. Error	0.197 (df = 239)	0.180 (df = 228)			185.162 (df = 239)	184.203 (df = 228)
F Statistic	0.543 (df = 3; 239)	4.415*** (df = 14; 228)			1.972 (df = 3; 239)	1.391 (df = 14; 228)
Note:						*p<0.05; **p<0.01; ***p<0.001

Figure 5:

```
## estimated average causal effect =  -0.001402144
## robust standard error =  0.03528588
## 95% confidence interval =  -0.07091319 0.0681089
## p-value =  0.9683362

##
## Model 2: covariate controls: ATE being in group CCT rather than group CCC:

## estimated average causal effect =  0.01283104
## robust standard error =  0.03313115
## 95% confidence interval =  -0.05245135 0.07811343
## p-value =  0.69891

##
## Model 1: simple: ATE being in group CTT rather than group CCC:

## estimated average causal effect =  -0.03923176
## robust standard error =  0.03436456
## 95% confidence interval =  -0.1069279 0.02846434
## p-value =  0.2547486

##
## Model 2: covariate controls: ATE being in group CTT rather than group CCC:

## estimated average causal effect =  -0.03299526
## robust standard error =  0.03178065
## 95% confidence interval =  -0.0956166 0.02962608
## p-value =  0.3002682

##
## Model 1: simple: ATE being in group TTT rather than group CCC:

## estimated average causal effect =  -0.02293274
## robust standard error =  0.03447528
## 95% confidence interval =  -0.09084694 0.04498147
## p-value =  0.506568

##
## Model 2: covariate controls: ATE being in group TTT rather than group CCC:

## estimated average causal effect =  -0.001795251
## robust standard error =  0.03009553
## 95% confidence interval =  -0.06109618 0.05750567
## p-value =  0.9524852
```

Results from logistic models 3 and 4 agrees with model 1 and 2, being assigned to any of the treatment groups rather than group CCC does not change the odds of achieving full accuracy to significantly higher or lower. The point estimates in model 3 suggests that being in any group rather than CCC actually decreases the odds of achieving full accuracy. The signs are flipped in model 4 for group CCT and TTT but the corresponding point estimates are very small in magnitude (no practical significance).

```
##
## Model 3: simple: ATE being in group CCT rather than group CCC:

##
## point estimate (linear combination) =  -0.009577293
## point estimate (average causal effect as odds ratio) =  0.9904684
## robust standard error (linear combination) =  0.2430156
## 95% confidence interval (linear combination) =  0.6136667 1.598633
## 95% confidence interval (average causal effect as odds ratio) =  0.6136667 1.598633
```

```
## p-value =  0.9685634

##
## Model 4: covariate controls: ATE being in group CCT rather than group CCC:

##
## point estimate (linear combination) =  0.08426371
## point estimate (average causal effect as odds ratio) =  1.087916
## robust standard error (linear combination) =  0.261319
## 95% confidence interval (linear combination) =  0.6500894 1.820612
## 95% confidence interval (average causal effect as odds ratio) =  0.6500894 1.820612
## p-value =  0.7471078

##
## Model 3: simple: ATE being in group CTT rather than group CCC:

##
## point estimate (linear combination) =  -0.2486255
## point estimate (average causal effect as odds ratio) =  0.7798719
## robust standard error (linear combination) =  0.2195383
## 95% confidence interval (linear combination) =  0.5060586 1.201838
## 95% confidence interval (average causal effect as odds ratio) =  0.5060586 1.201838
## p-value =  0.2574273

##
## Model 4: covariate controls: ATE being in group CTT rather than group CCC:

##
## point estimate (linear combination) =  -0.227606
## point estimate (average causal effect as odds ratio) =  0.796438
## robust standard error (linear combination) =  0.2178505
## 95% confidence interval (linear combination) =  0.5184749 1.223422
## 95% confidence interval (average causal effect as odds ratio) =  0.5184749 1.223422
## p-value =  0.2961242

##
## Model 3: simple: ATE being in group TTT rather than group CCC:

##
## point estimate (linear combination) =  -0.1498256
## point estimate (average causal effect as odds ratio) =  0.8608581
## robust standard error (linear combination) =  0.2264477
## 95% confidence interval (linear combination) =  0.5510587 1.344823
## 95% confidence interval (average causal effect as odds ratio) =  0.5510587 1.344823
## p-value =  0.5082054

##
## Model 4: covariate controls: ATE being in group TTT rather than group CCC:

##
## point estimate (linear combination) =  0.001111663
## point estimate (average causal effect as odds ratio) =  1.001112
## robust standard error (linear combination) =  0.2213911
## 95% confidence interval (linear combination) =  0.6471854 1.548592
## 95% confidence interval (average causal effect as odds ratio) =  0.6471854 1.548592
## p-value =  0.9959936
```

Results from linear models 5 and 6 suggest no significant effects on time spent for being assigned to any group other than group CCC.

```
##
## Model 5: simple: ATE being in group CCT rather than group CCC:
##
## estimated average causal effect =  21.94609
## robust standard error =  31.5166
## 95% confidence interval =  -40.1397 84.03188
## p-value =  0.4868961
##
## Model 6: covariate controls: ATE being in group CCT rather than group CCC:
##
## estimated average causal effect =  24.71263
## robust standard error =  34.61848
## 95% confidence interval =  -43.50041 92.92568
## p-value =  0.4760459
##
## Model 5: simple: ATE being in group CTT rather than group CCC:
##
## estimated average causal effect =  -1.231698
## robust standard error =  28.55516
## 95% confidence interval =  -57.48363 55.02023
## p-value =  0.9656307
##
## Model 6: covariate controls: ATE being in group CTT rather than group CCC:
##
## estimated average causal effect =  -5.638537
## robust standard error =  28.81673
## 95% confidence interval =  -62.41969 51.14262
## p-value =  0.8450436
##
## Model 5: simple: ATE being in group TTT rather than group CCC:
##
## estimated average causal effect =  68.96467
## robust standard error =  37.67417
## 95% confidence interval =  -5.251168 143.1805
## p-value =  0.06841212
##
## Model 6: covariate controls: ATE being in group TTT rather than group CCC:
##
## estimated average causal effect =  68.49357
## robust standard error =  38.11267
## 95% confidence interval =  -6.604519 143.5917
## p-value =  0.07363773
```

6.2.4.2 Group and Session Specific comparisons (Design 2 Main)

In this second approach, we take advantage of the group and session specific data the we have. This provides us more granularity for analysis. For example we can look at how outcomes changes from one session to another with or without treatment, how turkers' performance may build-up or deplete from one session to

another, or effect of receiving treatment in a specific session. This is our specified model in the form we specify and the form which R outputs. Mathematically they are exactly he same model. We separately estimated a version with covariates control, because the estimates remained the same, and the corresponding standard errors and p-values only changed marginally, we do not include it in this report in order to more clearly explain the mechanism of this estimation approach. The corresponding model and code reside on our repo.

(7) $logit(accuracy) = \beta_{intercept} + \beta_{GS} group * session$ (Compressed Form)

(7.1) $logit(accuracy) = \beta_1 + \beta_2 * groupCCT + \beta_3 * groupCTT + \beta_4 * groupTTT + \beta_5 * session2 + \beta_6 * session3 + \beta_7 groupCCT : session2 + \beta_8 groupCTT : session2 + \beta_9 groupTTT : session2 + \beta_{10} groupCCT * session3 + \beta_{11} groupCTT * session3 + \beta_{12} groupTTT * session3$ (Expanded Form)

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.8235294  0.0282108 29.1920   <2e-16 ***
## groupCCT               -0.0010142  0.0383125 -0.0265   0.9789
## groupCTT               -0.0431373  0.0401086 -1.0755   0.2825
## groupTTT               -0.0284926  0.0391139 -0.7285   0.4666
## roundtwo               -0.0173578  0.0369009 -0.4704   0.6382
## roundthree              0.0144648  0.0355561  0.4068   0.6843
## groupCCT:roundtwo      -0.0049546  0.0545827 -0.0908   0.9277
## groupCTT:roundtwo       0.0163774  0.0536613  0.3052   0.7603
## groupTTT:roundtwo       0.0072475  0.0536400  0.1351   0.8926
## groupCCT:roundthree     0.0037908  0.0534056  0.0710   0.9434
## groupCTT:roundthree    -0.0046609  0.0530039 -0.0879   0.9300
## groupTTT:roundthree     0.0094323  0.0524767  0.1797   0.8574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rboust Standard Errors are applied

From the regression output, we see that none of the coefficients we estimated are significant. Nevertheless, we proceed with the intended approach to demonstrate the exercise. Using the coefficients in the estimated form and our design graph, we can make between group, between session, difference-in-differences and overall comparisons. Figure 6 best explains this concept.

To arriving at the specific comparisons with estimated coefficients from the regression model, we derive each comparison as a combination of the estimated coefficients. The relationships are listed below.

Notation	Meaning	Linear Combination
CCC S2-S1	effect of being in round two rather than round one in group CCC	b_5
CCC S3-S1	effect of being in round three rather than round one in group CCC	b_6
CCC S3-S2	effect of being in round three rather than round two in group CCC	b_6 - b_5
CCT S2-S1	effect of being in round two rather than round one in group CCT	b_5 + b_7
CCT S3-S1	effect of being in round three rather than round one in group CCT	b_6 + b_10
CCT S3-S2	effect of being in round three rather than round two in group CCT	b_6 - b_5 + b_10 - b_7
CTT S2-S1	effect of being in round two rather than round one in group CTT	b_5 + b_8

Notation	Meaning	Linear Combination
CTT S3-S1	effect of being in round three rather than round one in group CTT	b_6 + b_11
CTT S3-S2	effect of being in round three rather than round two in group CTT	b_6 - b_5 + b_11 - b_8
TTT S2-S1	effect of being in round two rather than round one in group TTT	b_5 + b_9
TTT S3-S1	effect of being in round three rather than round one in group TTT	b_6 + b_12
TTT S3-S2	effect of being in round three rather than round two in group TTT	b_6 - b_5 + b_12 - b_9
T-C S1	effect of receiving treatment in Round one	b_4 - (b_3 + b_2)/3
T-C S2	effect of receiving treatment in Round two (regardless of round one status)	(-b_2 + b_3 + b_4 - b_7 + b_8 + b_9)/2
T-C S3	effect of receiving treatment in Round three (regardless of round one or two status)	(b_4 + b_12) - (b_2 + b_3 + b_10 + b_11)/3
TTT-CCC	effect of being in a group TTT that receive all three rounds of treatment rather than all control group CCC	b_4 + (b_9 + b_12)/3
(T)-CCC	effect of being in any treatment groups rather than all control group CCC	(b_2 + b_3 + b_4)/3 + (b_7 + b_8 + b_9 + b_10 + b_11 + b_12)/9
D_in_D S3	effect on Round 3, of Receiving Treatment in Round 3 but not earlier	b_10 - b_7
D_in_D S2	effect on Round 2, of Receiving Treatment in Round 2 but not earlier	- (1/2)*b_7 + b_8

Using general linear hypothesis, we estimate the point estimate, standard error and p-values for the above listed comparisons. As we have expected, none of the between-group, between-session, difference-in-differences, or overall comparisons are significant. Robust standard errors are applied.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = round_accuracy ~ group * round, data = by_Session.table)
##
## Linear Hypotheses:
##                   Estimate Std. Error t value Pr(>|t|)
## CCC S2-S1 == 0 -0.0173578  0.0369009  -0.470    1.000
## CCC S3-S1 == 0  0.0144648  0.0355561   0.407    1.000
## CCC S3-S2 == 0  0.0318226  0.0321593   0.990    0.988
## CCT S2-S1 == 0 -0.0223124  0.0402194  -0.555    1.000
## CCT S3-S1 == 0  0.0182556  0.0398488   0.458    1.000
## CCT S3-S2 == 0  0.0405680  0.0431454   0.940    0.992
## CTT S2-S1 == 0 -0.0009804  0.0389598  -0.025    1.000
## CTT S3-S1 == 0  0.0098039  0.0393088   0.249    1.000
## CTT S3-S2 == 0  0.0107843  0.0379124   0.284    1.000
## TTT S2-S1 == 0 -0.0101103  0.0389303  -0.260    1.000
## TTT S3-S1 == 0  0.0238971  0.0385949   0.619    1.000
## TTT S3-S2 == 0  0.0340074  0.0392051   0.867    0.996
## T-C S1 == 0    -0.0137755  0.0314239  -0.438    1.000
## T-C S2 == 0    -0.0210181  0.0273768  -0.768    0.999
```

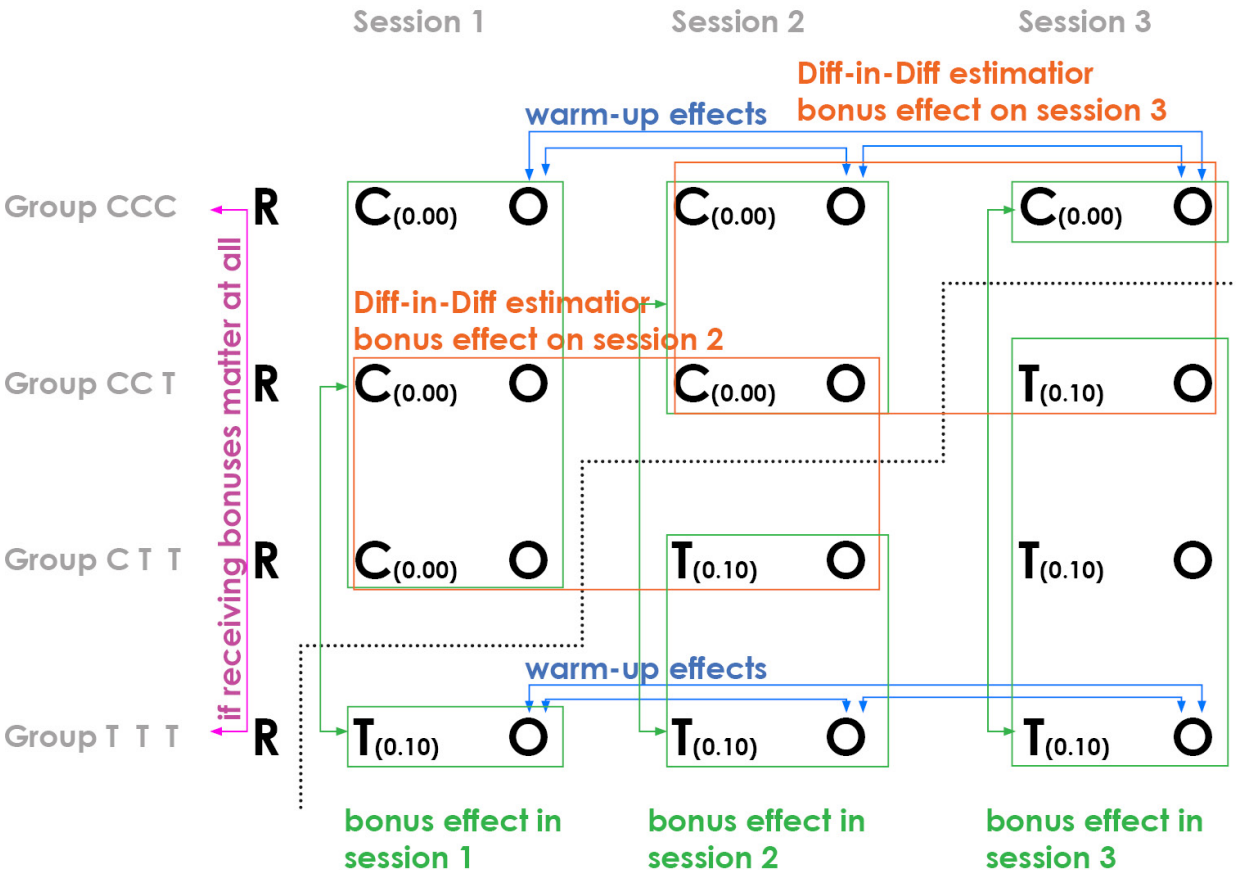

Session 1
Session 2
Session 3
Diff-in-Diff estimatior
bonus effect on session 3
warm-up effects
Group CCC
R
C(0.00)
O
C(0.00)
O
C(0.00)
O
Diff-in-Diff estimatior
bonus effect on session 2
Group CC T
R
C(0.00)
O
C(0.00)
O
T(0.10)
O
Group C T T
R
C(0.00)
O
T(0.10)
O
T(0.10)
O
if receiving bonuses matter at all
warm-up effects
Group T T T
R
T(0.10)
O
T(0.10)
O
T(0.10)
O
bonus effect in session 1
bonus effect in session 2
bonus effect in session 3

Figure 6:

```
## T-C S3 == 0    -0.0040532  0.0314754  -0.129    1.000
## TTT-CCC == 0   -0.0229327  0.0213467  -1.074    0.979
## (T)-CCC == 0   -0.0211889  0.0170422  -1.243    0.943
## D_in_D S3 == 0  0.0087454  0.0538121   0.163    1.000
## D_in_D S2 == 0  0.0188547  0.0475677   0.396    1.000
## (Adjusted p values reported -- single-step method)
```

Rboust Standard Errors are applied

6.2.4.3 Intermediate and Lagged Effects (Design 2 Main)

Finally, in this last approach, we want to derive a more summarizing effect of receiving bonus on accuracy for any one session. This exact mechanism is referenced from *Chapter 8.5 Field Experiment by Gerber and Green).To that end, we first state two substantive assumptions. Invoking the no anticipation assumption allow us to ignore treatments that are allocated after the current period. We also ignore treatments that were allocated two sessions prior, which is equivalent to assuming that bonuses' effects wask out entirely after two sessions. We redefine the potential outcomes as follow:

- **Y__00** (untreated during preceding and current session)
- **Y__01** (untreated during preceding session but treated during current session)
- **Y__11** (treated during preceding and current session)

The truncated tables below shows the randomly assigned treatment condition and observed outcomes for 5 of the 243 subjects under the new outcome definition.

```
## [1] "Treatment Condition Table:"
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	01	11	11
2	A10TT0QTE6FNQQ	01	11	11
3	A10XK0DWEXO5BY	00	01	11
4	A1131ZL7O5GM6R	00	01	11
5	A11QDNT3W7DT7K	00	00	01

```
## [1] "Observed Outcomes Table:"
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	0.5294118	0.1176471
2	A10TT0QTE6FNQQ	1.0000000	0.9411765	1.0000000
3	A10XK0DWEXO5BY	0.4705882	0.3529412	0.6470588
4	A1131ZL7O5GM6R	1.0000000	0.9411765	0.9411765
5	A11QDNT3W7DT7K	0.8235294	0.7058824	0.8235294

Note that we cannot naively compare average outcomes of treated sessions to the average outcomes of the untreated sessions. This approach is biased because it ignores the fact that the probability of assignment to treatment varies from session to session because individuals are much more likely to be assigned to bonuses in the final week than in the first week. It is also prone to biase because it ignores lagged effects, treating Y_11 and Y_01 as though they were identical.Instead we begin by calculating the probability of begin assigned to each treatment during each session. These probabilities are displayed below.

Treatment_Condition	Week_1	Week_2	Week_3
Pr(00)	0.75	0.50	0.25
Pr(01)	0.25	0.25	0.25

Treatment_Condition	Week_1	Week_2	Week_3
Pr(11)	0.00	0.25	0.50

Then, in order to obtain unbiased treatment effects, we apply inverse weights. Treated units are weighted by the inverse of the probability of being treated; control units are weighted by the inverse of the probability of being in control. According to the above probability table, calculation of the immediate effect is as follows: (programming code in .rmd file)

$$\hat{E}[Y_{01} - Y_{00}] = \frac{\frac{\sum_{S1} Y_{01}}{0.25} + \frac{\sum_{S2} Y_{01}}{0.25} + \frac{\sum_{S3} Y_{01}}{0.25}}{\frac{64}{0.25} + \frac{60}{0.25} + \frac{60}{0.25}} - \frac{\frac{\sum_{S1} Y_{00}}{0.75} + \frac{\sum_{S2} Y_{00}}{0.50} + \frac{\sum_{S3} Y_{00}}{0.25}}{\frac{179}{0.75} + \frac{119}{0.50} + \frac{61}{0.25}}$$

```
##
## Estimated Combined Immediate Effect: -0.02847309
```

In order to estimate effects for combined immediate and lag effect, we restrict our attention to the second and third sessions, because this type of treatment cannot occur in the first week. (programming code in .rmd file)

$$\hat{E}[Y_{11} - Y_{00}] = \frac{\frac{\sum_{S2} Y_{11}}{0.25} + \frac{\sum_{S3} Y_{11}}{0.50}}{\frac{64}{0.25} + \frac{124}{0.50}} - \frac{\frac{\sum_{S2} Y_{00}}{0.50} + \frac{\sum_{S3} Y_{00}}{0.25} +}{\frac{119}{0.50} + \frac{61}{0.25}}$$

```
##
## Estimated Combined Immediate and Lagged Effect: -0.02602659
```

Our point estimates are mildly negative, this result works against our research hypothesis and suggest that the session accuracy drops by a little below 3 percentage points if the turker was given a bonus for that session. These estimates are also very small in magnitude and may not be significant at all. To estimate the standard errors, we first look at the observed schedule of outcomes for this experiment. The tables below shows the schedule for the same five turkers.

```
## Observed Y00 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	NA	NA	NA
2	A10TT0QTE6FNQQ	NA	NA	NA
3	A10XK0DWEXO5BY	0.4705882	NA	NA
4	A1131ZL7O5GM6R	1.0000000	NA	NA
5	A11QDNT3W7DT7K	0.8235294	0.7058824	NA

```
## Observed Y01 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	NA	NA
2	A10TT0QTE6FNQQ	1.0000000	NA	NA
3	A10XK0DWEXO5BY	NA	0.3529412	NA
4	A1131ZL7O5GM6R	NA	0.9411765	NA
5	A11QDNT3W7DT7K	NA	NA	0.8235294

```
## Observed Y11 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	NA	0.5294118	0.1176471

V1	worker_id	one	two	three
2	A10TT0QTE6FNQQ	NA	0.9411765	1.0000000
3	A10XK0DWEXO5BY	NA	NA	0.6470588
4	A1131ZL7O5GM6R	NA	NA	0.9411765
5	A11QDNT3W7DT7K	NA	NA	NA

Then, we fill in the implied hypothetical schedule of potential outcomes under the assumption of constant treatment effects.

```
##
## Hypothetical Y01 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4990613	0.5554384	0.1436737
2	A10TT0QTE6FNQQ	1.0284731	0.9672031	1.0260266
3	A10XK0DWEXO5BY	0.4705882	0.3814143	0.6730854
4	A1131ZL7O5GM6R	1.0000000	0.9696496	0.9672031
5	A11QDNT3W7DT7K	0.8235294	0.7058824	0.8520025

```
##
## Hypothetical Y01 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	0.5269653	0.1152006
2	A10TT0QTE6FNQQ	1.0000000	0.9387300	0.9975535
3	A10XK0DWEXO5BY	0.4421151	0.3529412	0.6446123
4	A1131ZL7O5GM6R	0.9715269	0.9411765	0.9387300
5	A11QDNT3W7DT7K	0.7950563	0.6774093	0.8235294

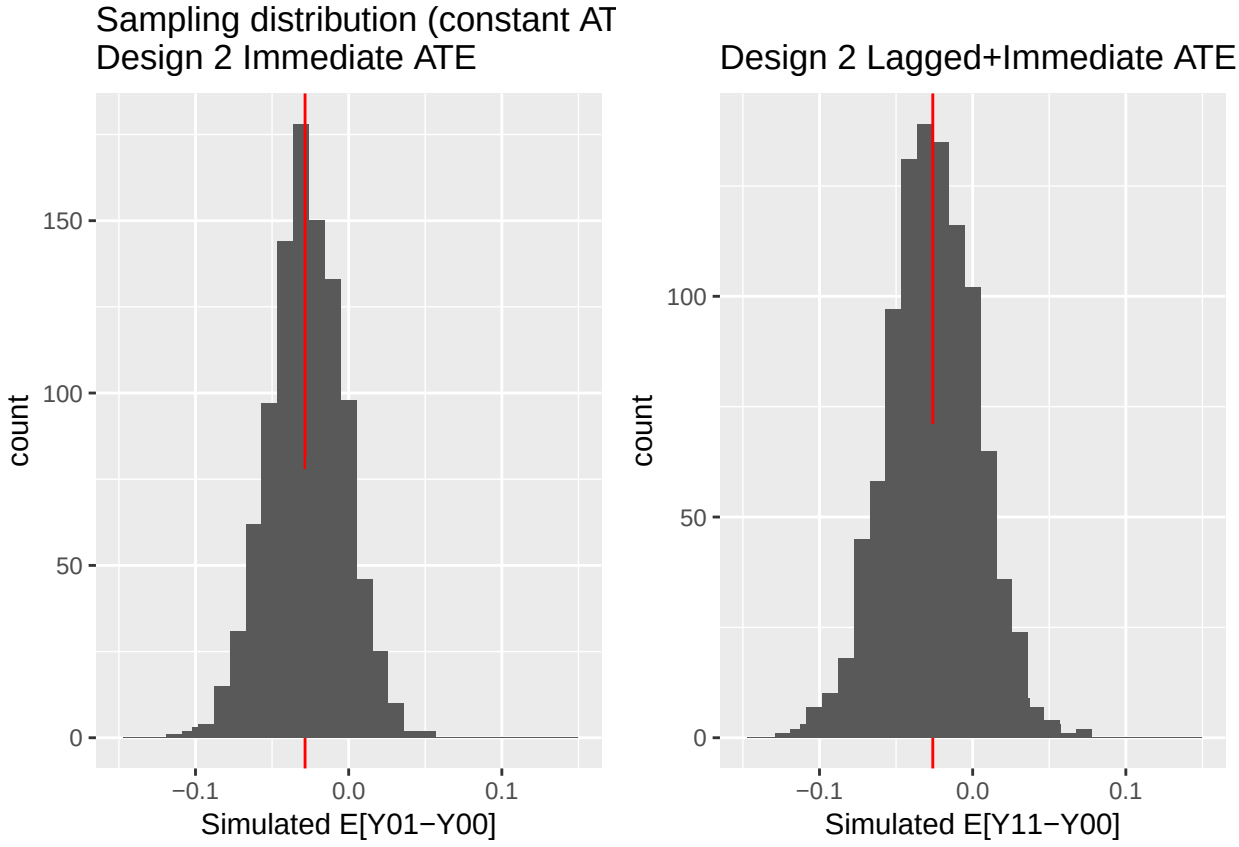
```
##
## Hypothetical Y01 Table:
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4730347	0.5294118	0.1176471
2	A10TT0QTE6FNQQ	1.0024465	0.9411765	1.0000000
3	A10XK0DWEXO5BY	0.4445616	0.3553877	0.6470588
4	A1131ZL7O5GM6R	0.9739734	0.9436230	0.9411765
5	A11QDNT3W7DT7K	0.7975028	0.6798558	0.8259759

Finally, we simulate the sampling distribution by shuffling the treatment group assignment the turkers, keeping the number of turkers originally assigned to each group the same. With the sampling distributions, we take the 0.05 and 0.975 quantiles to construct our confidence intervals. As we have expected, neither immediate treatment effect nor combined immediate treatment effect is statistically significant. The associated confidence intervals both cover zero. (programming code in .rmd file and repo)

```
##
## First 50 simulated values of E[Y01-Y00]--immediate treatment effect:

##  [1] -0.0349393629 -0.0284045117 -0.0742562974 -0.0661612363  0.0027973061
##  [6] -0.0430953146 -0.0126777127 -0.0339075717 -0.0268015509 -0.0142133201
```

```
## [11] -0.0324361580 -0.0164952355 -0.0232226267 -0.0491450446 -0.0120617614
## [16] -0.0014931410 -0.0125166520 -0.0392402424 -0.0319151788 -0.0379062796
## [21] -0.0278358946 -0.0098454358 -0.0129723361 -0.0343048415 -0.0017680101
## [26] -0.0585111347 -0.0425619773  0.0013252044 -0.0508515972 -0.0393183560
## [31]  0.0062130811 -0.0093178786 -0.0360478246  0.0112798185 -0.0513954762
## [36] -0.0285862617  0.0268823662 -0.0016686261 -0.0050020272 -0.0190551465
## [41] -0.0130932387 -0.0004110031 -0.0151512976  0.0055490819 -0.0405638652
## [46] -0.0324560857 -0.0364143395 -0.0762825048 -0.0221137960 -0.0269029508

##
## First 50 simulated values of E[Y11-Y00]--combined immediate and lagged treatment effect:

##  [1] -0.0444730982 -0.0360289136 -0.0684685342 -0.0648542352  0.0167091116
##  [6] -0.0120527916 -0.0033238781 -0.0461331720 -0.0055368960 -0.0534630696
## [11] -0.0373633033 -0.0166959118 -0.0141919509 -0.0347646151 -0.0383833325
## [16] -0.0047355887 -0.0355714495 -0.0403696359 -0.0449658003 -0.0697993524
## [21]  0.0013410633 -0.0379908546 -0.0255325936 -0.0197994136 -0.0046603591
## [26] -0.0724522894 -0.0510419319  0.0018912487 -0.0372710683 -0.0321502683
## [31]  0.0188205053 -0.0220823469 -0.0581317942 -0.0106028570 -0.0255298848
## [36] -0.0435323075  0.0394452371  0.0066753804 -0.0050138653 -0.0105440527
## [41] -0.0351383545 -0.0003346431 -0.0122667931  0.0290164898 -0.0609990731
## [46] -0.0087089908 -0.0169425313 -0.0750009689 -0.0206832068 -0.0277418634

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




```
##
## Estimated Confidence Interval of E[Y01-Y00]--immediate treatment effect:
```

```
##        2.5%       97.5%
## -0.07630711  0.02144866
```

```
##
## Estimated Confidence Interval of E[Y11-Y00]--combined immediate and lagged treatment effect:
```

```
##        2.5%       97.5%
## -0.08285649  0.02994102
```

6.3 Design 2 conclusion

To provide an overview for the design 2 results, the estimated effect of bonuses on a turker performance is statistically and practically insignificant in all three of our estimation approaches. We looked at the overall bonus effects by comparing different treatment groups using linear and logistic regression, investigated more granular bonus effects both by group and by session by taking full advantage of a stepped-wedge design, and we estimated both immediate, lagged and their combined effect by invoking the no-anticipation assumption and assuming no within-subject spill-over effect from more than one session. All inference approaches failed to reject the null hypothesis that bonus has zero effect on a turker's performance. No significant effects were estimated when time spent or screeners are studied as outcomes. Finally, we observed differential attrition based on covariates related to turkers' prior exposure or life experience with dogs. This remains a threat to the unbiasedness of our estimated treatment effect. We will discuss potential solutions for a future experiment in section 7.

7 Limitations and Future Lines of Investigation

As mentioned in our design 1 conclusion, something related to our publishing orders affected the outcomes. It is unlikely but possible that the issue is related to the weekend/time of day the experiment was conducted, as well as (but unlikely) running out of turks after our first publish order. To improve on these issues, we would suggest an extension of the experiment schedule table found in page 4 of this report, with at least two more publish orderings, to fully randomize our treatment by day of the weekend and time of day. We would also like to stay away from long holiday weekends; our first ordering occurred on Veteran's day weekend and this may have been or contributed to the discrepancy in the $0.10 treatment we saw between the orderings. These additional rounds may also increase our sample size.

A common way to mitigate threat of attrition for design 2, is to perform a second round of experiment which randomly re-sample the attriters. That is, re-invite those turkers to complete the HIT they started. Through the Qualtrics platform, we keep a list of turkers' identification numbers from those who did not complete their HITs. We will assign them the exact same base rate, same bonus conditions and same survey form over another weekend. However, we can only recover as many as possible and full re-participation is not guaranteed. Once completed, we will attach their data back on our existing design 2 data and conduct the same inference procedure to detect any differences in our estimations.

Regarding generalizability, we have different concerns for the two experimental setups. First, in design 1, because of the abnormalities we observed from the data collected over the veteran's day weekend, we are not certain if the corresponding results can be generalized to other weekends in general. Second, in design 2, because we assign treatment as bonuses, we are not certain if the same results can apply to assigning base rates directly. In each experimental design, we attempted four different pay rates bounded by $0.10 and $0.55. If the upper bound can be expanded to $0.80 the result may be different since turkers may fear higher chances of getting their work rejected by the requestors. Overall, we are not certain if an experiment conducted on the AMT platform can generalize to the labor market at large, because at the pay rates we see on the AMT platform, the full-time income cannot even add up to the minimum wage in California ($10.75 per hour in December 2017). Lastly, our task design is a very specific image classification task related to dog breeds. There are a much wider variety of tasks on the AMT platform which our task cannot represent. Needless to say, in the real world the labor market is full of demands and supplies of a much, much wider range of skills.

8 Conclusion

To conclude the project as a whole, the two experimental designs we attempted both estimated insignificant treatment effect related to payment. Design 1 looks at whether higher pay rate as published on the AMT platform increases quality of returned HIT for employers, and design 2 looks at whether assigning bonuses within each HIT provides more incentives for turkers to perform better. The two design2 tackle our research question – whether higher rewards have an effect on labor productivity, from similar but different perspectives. We discussed ways to improve the short-coming of each of our experimental designs, they mostly involve extending the experiment schedule to create more clusters of observations for our traditional between-subject design, and collect more responses from prior attriters to mitigate the threat of attrition for our stepped-wedge design. We also expressed doubts on the generalizability of our project, because the mechanism of the AMT platform and nature of our designed task are very specific. Results may not generalize well.

Finally, we find our results in-line with prior research projects mentioned in section 2, which suggested that higher financial incentives increases the quantity, but not quality of the work done by workers. In executing both of our experimental designs, we observed that for HIT postings attached with similar quantity of available HITs, those that are posted at a higher reward rate on the AMT platform tend to take much shorted time to finish. In addition, we didn't estimate any significant monetary effect on labor performance or returned task quality.
