BA830 Business Experiment

Christian Lawrence, Jiazheng Li, Michelle Lensing, Tiam Moradi, and Ying Zhang

Introduction

With remote learning becoming the norm across universities in the United States, teachers have had to innovate around the way they assess their students. Some universities and professors have adopted strict remote exam guidelines while others have taken advantage of the flexible situation and provided more relaxed guidelines, which begs the question(s): what's the best way to distribute a remote assessment? Does having a quiz with camera on affect a student's performance? Will having people sit on a call improve or worsen student scores?

This experiment aims to take a first pass at assessing the optimal method of administering remote assessments.

Hypothesis

Our team is split on whether the treatment or control would lead to better outcomes (higher scores and less time taken). One on hand, removing time constraints and the need for cameras may relieve pressure and allow students to perform better. On the other, students may take tests more seriously with the presence of a remote proctor and their cameras on. We also suspect that certain participant characteristics might affect the treatment effects. For example, English native speakers might score better than those who are not.

Experiment Design

Preliminary Survey

To conduct appropriate blocking randomization, we first gave the participants a survey to uncover demographic information. This information will be used for either blocking randomization or as covariates in our analysis. Additionally, we included a question around which time/date they preferred to take the assessment, providing four options across three days; their responses were used for scheduling purposes.

Treatment (N=61, treatment group = 32, control group = 29)

Control → Take assessment unproctored

Treatment → Take proctored assessment on a Zoom call with a camera on

Assessment and Delivery Method

The assessment consisted of 21 logic reasoning questions, which involved some math and some reading comprehension.

Participants on the Zoom call were given the assessment when they signed on to ensure everyone started the assessment at the same time. We did not provide a time limit for the assessment, so participants could take as long as they wanted; however, we did inform them that it takes an average of 10-15 minutes to complete.

The assessment was administered through Qualtrics. We modified the original questions and used images (instead of text) to display quiz questions on the Qualtrics platform to prevent participants from googling

answers. We also required participants to refrain from searching for the answers online or collaborating with other people during the assessment, simulating a closed book testing environment.

We scheduled participants to take the assessment on the time/date they selected on the preliminary survey, either sending them a Zoom link 5-10 minutes before their scheduled time and the assessment link during the call if they were in the treatment group, or sending them the assessment link at their scheduled time if they were in the control group.

Both groups received email reminders prior to the assessment, indicating whether they will be taking the assessment on Zoom or not. They were unaware of the existence of the alternative group (e.g. if they took the assessment proctored on Zoom, they were not informed of the unproctored group).

Outcomes

Our primary outcome measure will be how well the participants scored in the exam (i.e. how many questions they answered correctly). The second will be the time taken to complete the assessment. Although not our main outcome, we also did look into number of clicks as an outcome for an analysis.

Blocking Randomization

To make sure we have an even distribution of individuals with certain characteristics in both groups, we attempted to block randomize for the participants' education level, gender, and whether or not English was their first language.

Data

The quantitative variables in our dataset have all been converted to numerical variables in order for us to better analyze the results. The 'genderMale' variable is a 1 for a male and 0 for female. The age variable takes into account four age groups, 20-25, 25-30, 30-35, and 35+, these are 1, 2, 3, and 4, respectively. The 'english1' variable is a 1 if the participant's first language is English and a 0 otherwise. Lastly, the 'treatment' variable is a 1 for those in the treatment (on the Zoom call) and 0 for the control group.

```
quiz <- fread("/Users/michellelensing/Documents/MSBA/BA830 Bus Exp and Causal Methods/FinalExperiment.cs
head(quiz)
```

```
##           ResponseId Quiz_Timer Click score genderMale age education english1
## 1: R_3qdFrptt2liUU0n    681.973    28    12          1   1         2        0
## 2: R_2wnLT1jWElzy9w9   1710.179    39    14          1   1         2        0
## 3: R_2w5d2aO0nge9Pc6    113.755    22    13          1   3         2        0
## 4: R_e5S5Gh8REeeJiwN    827.600    28    14          1   1         2        1
## 5:  R_27p0gUnn5hDZ2O   1551.289    27    16          0   1         2        0
## 6: R_1dn6NNSfx5CWvOs    700.334    75    11          1   2         1        1
##    treatment
## 1:         1
## 2:         0
## 3:         0
## 4:         0
## 5:         0
## 6:         0
```

Pre-Experiment Characteristics

```
# Checking if the treatment and control have similar pre-experiment characteristics
quiz[treatment==1,mean(genderMale)] - quiz[treatment==0,mean(genderMale)]
```

```
## [1] -0.0549569
```

```
quiz[treatment==1,mean(education)] - quiz[treatment==0,mean(education)]
```

```
## [1] -0.07435345
```

```
quiz[treatment==1,mean(english1)] - quiz[treatment==0,mean(english1)]
```

```
## [1] 0.1584052
```

```
quiz[treatment==1,mean(age)] - quiz[treatment==0,mean(age)]
```

```
## [1] -0.07650862
```

We can see that for each of these conditions (gender, education, English first language, and age), our treatment and control groups do have very similar pre-experiment characteristics, which is a good indicator of proper randomization.

Check for Randomization and Blocking

```
num_treated = quiz[treatment ==1,.N]
total_obs = dim(quiz)[1]

prop.test(num_treated,total_obs,.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  num_treated out of total_obs, null probability 0.5
## X-squared = 0.065574, df = 1, p-value = 0.7979
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3937835 0.6522984
## sample estimates:
##         p
## 0.5245902
```

Proportions between Treatment and Control Groups: Based on the following proper randomization test, we fail to reject the null hypothesis, since the p-value is greater than 0.05 and our proportion of 0.5 lies within the confidence interval. The proportion of units treated is what we expected.

```
english = feols(english1 ~ treatment ,data=quiz)
ed = feols(education ~ treatment,data=quiz)
age = feols(age ~ treatment,data=quiz)
gender = feols(genderMale ~ treatment,data=quiz)

summary(english)
```

Blocking/Randomization

```
## OLS estimation, Dep. Var.: english1
## Observations: 61
## Standard-errors: Standard
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.310345   0.091022  3.4096 0.001179 **
```

```
## treatment   0.158405   0.125671  1.2605 0.212462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.482066   Adj. R2: 0.009718
```

```
summary(ed)
```

```
## OLS estimation, Dep. Var.: education
## Observations: 61
## Standard-errors: Standard
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)  1.793100   0.106007 16.915000 < 2.2e-16 ***
## treatment   -0.074353   0.146361 -0.508015  0.613337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.561429   Adj. R2: -0.01252
```

```
summary(age)
```

```
## OLS estimation, Dep. Var.: age
## Observations: 61
## Standard-errors: Standard
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  1.482800   0.150899  9.826200 4.94000e-14 ***
## treatment   -0.076509   0.208341 -0.367228 7.14763e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.799181   Adj. R2: -0.01463
```

```
summary(gender)
```

```
## OLS estimation, Dep. Var.: genderMale
## Observations: 61
## Standard-errors: Standard
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  0.586207   0.093641  6.260100 4.77000e-08 ***
## treatment   -0.054957   0.129288 -0.425074 6.72329e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.495938   Adj. R2: -0.013844
```

Interpretation: The regressions for all of these variables have high p-values (above 0.05) therefore, we fail to reject the null hypothesis and conclude that proper blocking and randomization is in effect.

Randomization Check for the Newly Created Age Groups

Most of the participants in our experiment are MSBA program students, whose ages are between 20-25, and others are friends/family of the team. To make the age distribution more balanced in our experiment analysis, we labeled the ages between 20-25 years old as 0, and labeled the ages above 25 years old as 1. We are interested to know whether the effect of taking a quiz in a Zoom meeting is different for people who are older or younger.

```
age_df<- copy(quiz)
#re-label the age level: 20-25 yrs old as 0, and above 25 yrs old as 1
age_df[,age_binary:=ifelse(age>1,1,0)]
t.test(age_df[treatment==1,age_binary],age_df[treatment==0,age_binary])
```

```
##
##  Welch Two Sample t-test
##
## data:  age_df[treatment == 1, age_binary] and age_df[treatment == 0, age_binary]
## t = -0.24446, df = 58.015, p-value = 0.8077
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2673305  0.2091409
## sample estimates:
## mean of x mean of y
## 0.2812500 0.3103448
```

Interpretation: The t.test analysis for the new age groups indicates a high p-value, so we fail to reject the null hypothesis and conclude that proper blocking and randomization is in effect for this variable as well. This indicates that the treatment and control groups have similar age levels.

Regression: Score on Treatment and Quiz Time on Treatment

```
reg_score <- feols(score ~ treatment , data = quiz, se = 'white')
reg_time <- feols(Quiz_Timer ~ treatment , data = quiz, se = 'white')

dict = c("treatment" = "Zoom_treatment")
etable(reg_score,reg_time, dict = dict)
```

```
##                          reg_score         reg_time
## Dependent Var.:              score       Quiz_Timer
##
## (Intercept)     15.31*** (0.4354) 912.8*** (76.74)
## Zoom_treatment    0.9397 (0.5909)   -57.37 (93.41)
## _______________ _________________ _________________
## S.E. type       Heteroskeda.-rob. Heterosked.-rob.
## Observations                   61                61
## R2                        0.04125           0.00655
## Adj. R2                   0.02500          -0.01029
```

Interpretation: According to the regression results above, treatment does not cause a statistically significant difference for score or quiz time. Taking quiz on Zoom might increase the participant's score by 0.94 points since the average treatment effect for score is 0.9397, but this effect is small and not statistically significant. Additionally, taking a quiz on Zoom also seems to reduce the time respondents spent on the quiz by 57 seconds, but the difference is not statistically significant.

Regression: Number of Clicks on Treatment

```
reg_click <-  feols(Click ~ treatment , data = quiz, se = 'white')

dict = c("treatment" = "Zoom_treatment")
etable(reg_click, dict = dict)
```

```
##                        reg_click
## Dependent Var.:            Click
##
## (Intercept)     48.03*** (5.790)
```

```
## Zoom_treatment   -13.57* (6.523)
## _______________ _________________
## S.E. type       Heterosked.-rob.
## Observations                  61
## R2                       0.07178
## Adj. R2                  0.05605
```

Interpretation: Surprisingly, the treatment group has much fewer clicks than the control group. According to the average treatment effect of -13.57 clicks, this difference is statistically significant. This could show that by taking the quiz over Zoom, participants felt more confident in their initial answers to questions, whereas participants who did not take the quiz over Zoom were more hesitant in their initial answers, feeling the need to change their answer a few times before moving on. A separate possible reasoning for this difference could be that taking a quiz over Zoom may cause people to feel more pressure to go faster, so maybe they will just stick with their first choice and move on.

Regression: Score on Treatment and Quiz Time on Treatment (with Covariates)

```
reg_score_cov <- feols(score ~ treatment + genderMale +
                         education + english1 , data = quiz, se = 'white')
reg_time_cov <-  feols(Quiz_Timer ~ treatment + genderMale +
                         education + english1, data = quiz, se = 'white')

dict = c("treatment" = "Zoom_treatment", "english1"="English_first_langague")
etable(reg_score_cov,reg_time_cov, dict = dict)
```

```
##                            reg_score_cov     reg_time_cov
## Dependent Var.:                    score       Quiz_Timer
##
## (Intercept)             15.23*** (1.318) 936.5*** (163.8)
## Zoom_treatment           0.7347 (0.6181)   -53.53 (98.59)
## genderMale              -0.5095 (0.6245)   -75.08 (105.4)
## education                0.0158 (0.6067)    18.56 (69.02)
## English_first_langague   1.124. (0.6609)   -41.62 (103.0)
## ______________________ _________________ _________________
## S.E. type               Heterosked.-rob. Heterosked.-rob.
## Observations                          61               61
## R2                               0.09796          0.02445
## Adj. R2                          0.03353         -0.04523
```

Interpretation: Prior to the experiment, we believed features like education level, whether English is first language or not, gender, and age, might affect quiz scores and quiz time, so we added these variables into our regression models as covariates.

However, the regression results indicate that there is no statistically significant difference for both score and quiz time on treatment, although we controlled these aforementioned factors. Additionally, the standard errors for the treatment actually increased slightly when adding in these covariates. This could indicate that the covariates are not correlated with the outcome, so these are just adding more noise.

Heterogeneity

CATE for English as first language

```
# running a regression where English language as first language
# is the interaction term and score is the outcome
reg_score_int <- feols(score ~ treatment* english1 , data = quiz, se = 'white')

dict = c("treatment" = "Zoom_treatment", "english1"="English_first_langague")
etable(reg_score_int, dict = dict)
```

```
##                                          reg_score_int
## Dependent Var.:                                  score
##
## (Intercept)                             15.55*** (0.5248)
## Zoom_treatment                           -0.4324 (0.7457)
## English_first_langague                   -0.7722 (0.9507)
## Zoom_treatment x English_first_langague   3.188** (1.164)
## _____________________________________ _________________
## S.E. type                               Heteroskeda.-rob.
## Observations                                           61
## R2                                                0.19542
## Adj. R2                                           0.15307
```

```
#CATE for English native speakers:
3.188 - 0.4324
```

```
## [1] 2.7556
```

```
#CATE for non-English native speakers:
-0.4324
```

```
## [1] -0.4324
```

The average treatment effect for English native speakers is 2.76 points. The average treatment effect for non-English native speakers is -0.4324 points.

Interpretation: We thought that English as a first language or not possibly could play a role when interacting with treatment. So, we decided to measure the average treatment effect conditioning on the variable "english1" (whether English is the participant's first language).

This led to an interesting finding– for English native speakers, taking a quiz in Zoom could increase their score by 2.765, while for a non-English native speaker, taking a quiz on Zoom might not help their performance or even slightly bring down their score by 0.43. This difference is statistically significant.

One possible explanation is cultural differences. English native speakers might feel more comfortable about taking quizzes in a Zoom meeting with camera on, compared to a non-English native speaker. However, we may need further analysis or an additional experiment with a larger sample size to determine whether this heterogeneous treatment effect exists.

Note: We are aware of the fact that the sample size is small, so we believe that further analysis and an additional experiment with a large sample size would be helpful to determine whether the language heterogeneity effects exist.

CATE for Age

```
#running a regression where age is the interaction term and quiz time is the outcome
reg_time_age <-  feols(Quiz_Timer ~ treatment* age_binary , data = age_df, se = 'white')
dict = c("treatment" = "Zoom_treatment", "age_binary"="age_above_25")
etable(reg_time_age,dict = dict)
```

```
##                                    reg_time_age
## Dependent Var.:                       Quiz_Timer
##
## (Intercept)                    1,047.5*** (92.60)
## Zoom_treatment                    -216.9. (110.7)
## age_above_25                     -434.0*** (118.6)
## Zoom_treatment x age_above_25     522.2** (173.8)
## ____________________________ __________________
## S.E. type                     Heteroskedas.-rob.
## Observations                                  61
## R2                                       0.16595
## Adj. R2                                  0.12205
```

```
#CATE for those who are above 25 (age==1):
-216.9 + 522.2
```

```
## [1] 305.3
```

```
#CATE for those who are between 20-25 (age==0):
-216.9
```

```
## [1] -216.9
```

The average treatment effect for those who are older than 25 years old is 305.3 seconds when quiz time is the outcome. The average treatment effect for those who are between 20 to 25 years old is -216.9 seconds when quiz time is the outcome.

Interpretation: These differences are statistically significant. We found that for those who are above 25 years old, taking a quiz on Zoom increases their time spent on the quiz by around 305 seconds, which is about 5 minutes. In comparison, for those who are 20-25 years old, taking a quiz on Zoom decreases their quiz time by 217 seconds, which is about 3.6 minutes.

One possible explanation is that most of the 20-25 age group participants are MSBA students who are used to the Zoom environment. In comparison, for those who are 25+ years old, they might not be used to taking a quiz in a Zoom meeting and any technical issues might take more time for them to complete the quiz.

Note: We are aware of the fact that the sample size is small and relatively fewer people who are above 25 years old in our experiment. We believe further analysis and an additional experiment with a large sample size would be helpful to determine whether the age heterogeneity effects exist.

Appendix:

```
#running a regression where gender is the interaction term and score is the outcome
reg_gender_score <-feols(score ~ treatment*genderMale, data = quiz, se= 'white')
etable(reg_gender_score)
```

```
##                          reg_gender_score
## Dependent Var.:                     score
##
## (Intercept)             15.75*** (0.6580)
## treatment                 0.4500 (0.8598)
## genderMale               -0.7500 (0.8795)
## treatment x genderMale    0.8441 (1.194)
## ______________________ _________________
## S.E. type              Heteroskeda.-rob.
## Observations                          61
```

```
## R2                                0.05362
## Adj. R2                           0.00381
```

```
#CATE for Males:
0.84 + 0.45
```

```
## [1] 1.29
```

```
#CATE for Females:
0.45
```

```
## [1] 0.45
```

The average treatment effect for males is 1.29. The average treatment effect for females is 0.45.

Interpretation: We wanted to see whether gender played a role on score, when interacting with treatment. We can see here that while both males and females get a slightly higher score on average when taking the quiz over Zoom versus not, male scores increase by more than female scores. Male scores will increase by about 1.29 points when taking the quiz over Zoom versus on their own, while female scores will only increase 0.45 points when taking the quiz over Zoom, however, this difference is not statistically significant.

```
#running a regression where age is the interaction term and score is the outcome
reg_score_int <- feols(score ~ treatment* age , data = quiz, se = 'white')
etable(reg_score_int)
```

```
##                     reg_score_int
## Dependent Var.:             score
##
## (Intercept)     15.80*** (0.7275)
## treatment         -0.1020 (1.050)
## age               -0.3297 (0.4490)
## treatment x age    0.7228 (0.5666)
## _______________ _________________
## S.E. type       Heteroskeda.-rob.
## Observations                   61
## R2                        0.05703
## Adj. R2                   0.00740
```

```
#CATE for those who are older:
0.7228 - 0.1020
```

```
## [1] 0.6208
```

```
#CATE for those who are younger:
-0.1020
```

```
## [1] -0.102
```

The average treatment effect for those who are older is 0.621 when score is the outcome. The average treatment effect for those who are younger is -0.102 when score is the outcome.

Interpretation: We were interested to see whether age played a role on score, when interacting with treatment. We found that for those that are older, taking a quiz on Zoom increases their score by about 0.62 points, whereas for those who are younger, their quiz scores decrease by 0.1 points. However, this is not statistically significant.

```
#running a regression where English as a first language is the interaction term and quiz time is the ou
reg_time_int <-  feols(Quiz_Timer ~ treatment* english1, data = quiz, se = 'white')
etable(reg_time_int)
```

```
##                           reg_time_int
## Dependent Var.:             Quiz_Timer
##
## (Intercept)           911.9*** (93.26)
## treatment               -3.213 (123.3)
## english1                 2.836 (170.4)
## treatment x english1    -116.5 (200.2)
## ____________________ _________________
## S.E. type             Heterosked.-rob.
## Observations                        61
## R2                             0.02001
## Adj. R2                       -0.03157
```

```
#CACE for English native speakers:
-3.213 - 116.5
```

```
## [1] -119.713
```

```
#CACE for non-English native speakers:
-116.5
```

```
## [1] -116.5
```

The average treatment effect for English native speakers is -119.7 seconds. The average treatment effect for non-English native speakers is -116.5 seconds.

Interpretation: We also wanted to see whether English as a first language played a role in the quiz time, when interacting with treatment. Here we can see that for those who are native English speakers, taking a quiz over Zoom allows them to complete it 120 seconds (or 2 minutes) faster than when taking it not over Zoom. For non-English native speakers, this time decrease is a few seconds less (116 seconds) when taking the quiz over Zoom versus not over Zoom. However, these differences are not statistically significant.