Effect of Multiple Attempts Assumption On Quiz Scores

Team 13 Antonio Moral Cevallos Bosoo Kim Jiajian(Sylar) Guo
Manushi Patel Ying(Amber) Wu Yixuan Wang

Introduction

With the development of technology, more and more tests and exams are taken online which provides students a bigger platform to utilize and better assist quality study. In some circumstances, students are allowed to have extra attempts on their assignments and tests to raise the quality of the work. Admittedly, more attempts usually lead to higher scores in most of the cases. Not only did Professor Luebben mentioned in her article that in the experiment she conducted in her class students improved their scores by an average of 0.71 points with multiple attempts, Dr. Archer also mentioned in her research paper that the mean exam score increased from 60.34% to 70.77% when multiple attempts were allowed.

Knowing that multiple attempts could potentially lead to higher scores in the end, do people tend to get lower scores in their initial attempt when they are given multiple attempts because only the last attempt counts and they think they will have extra chances to improve the answers? Our team decided to conduct an experiment by distributing surveys containing six cognitive test questions to two different groups. People in the treatment group are allowed to answer these questions with multiple attempts and they will see their scores after each attempt. In this experiment, our alternative hypothesis is that people tend to get lower scores in their initial attempt if they are given multiple attempts. In other words, the mean initial score in the treatment group is lower than the mean score in the control group. Our null hypothesis is that the mean initial score among the treatment group has no difference from the mean score among the control group.

```
# Read the data
treatment <- fread('Cognitive_Test_1.csv')
control<- fread('Cognitive_Test_2.csv')
treatment <- treatment[, treatment := 1]
control <- control[, treatment := 0]
total <- rbind(treatment, control)
```

Method

To successfully conduct the experiment, our team established structured survey questions and distributed them to our target population, 68 MSBA students, excluding 6 project members. The main communication channel was email. We opened the surveys for 7 days, and we also received the survey results to see how participants responded and to make sure we received correct data from the participants. One week after we sent our initial survey, we noticed that only 27 out of 68 people answered our survey, so we contacted individuals through Slack Channel (Communication Platform) to encourage the participation. Two days after we sent a reminder message to participants, we closed our experiment. This experiment took a total of 9 days, and 66% of the survey population or 45 people responded to the survey, 19 treatment groups and 26 control groups, respectively.

Treatment

We decided our target people would be the students of the MSBA to reduce the bias in our sample since we would have similar cognitive abilities. We surveyed 68 students from the MSBA group, so there were 34 people in the treatment and the control. For our treatment group we decided to add one line before

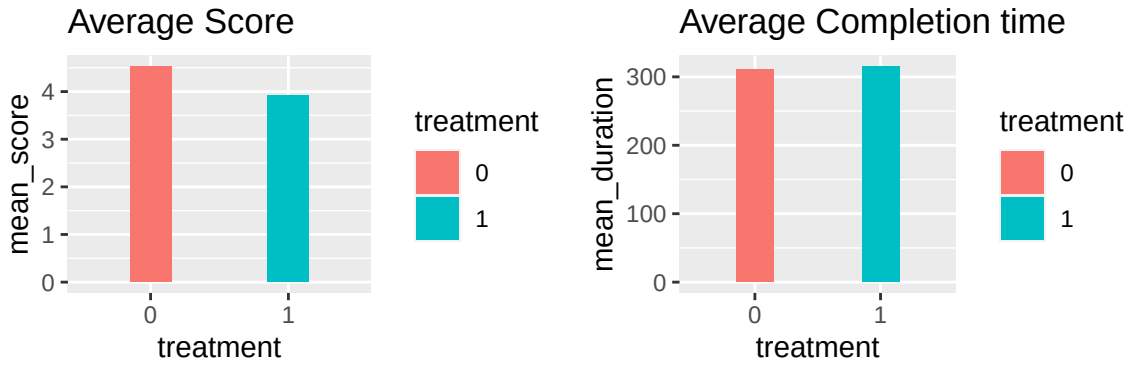
the survey page: 'You will have multiple attempts to complete the questions.' (but only consider their first attempt of the questions)

Randomization

We-randomized in our MSBA group ie we divided 50% of people into treatment and 50% into the control. One of the problems we faced in doing so is that we did not control for the gender while dividing our subjects and that came up in our analysis. We could have solved this problem by using the block randomization technique where we control for the proportion of the gender in the particular groups.

Data Analysis

Firstly, we did two plots to see the difference between treatment group and control group on average score and average completion time. We found that average score of control group is slightly higher than that of treatment group, but average completion time of treatment group is slightly higher than that of control group.



t-test Since p-value is larger than 0.05, the true difference in mean scores between the control group and treatment group is not statistically significant from 0. In other words, the mean scores between two groups are not significantly different. Similarly, the mean completion time between two groups are also not significantly different.

```
t1 <- t.test(total[treatment == 1, score], total[treatment == 0, score])
t2 <- t.test(total[treatment == 1, duration], total[treatment == 0, duration])
map_df(list(t1, t2), tidy)
```

```
## # A tibble: 2 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
## 1   -0.603      3.92      4.53   -1.42     0.162      43.0    -1.46     0.252
## 2    4.64     315.      310.      0.0956   0.924      42.5   -93.3    103.
## # ... with 2 more variables: method <chr>, alternative <chr>
```

Randomization check From our randomization analysis we learned that our randomization was partially effective. Our treatment and control groups were statistically similar when it came to their ages and their work experience. However, when looking at the distribution of gender and GPA, we see that the groups in our experiment are statistically different. Seeing these results led to some analysis to understand why these values could be so different in these aspects. Given that for the entire sample (treatment and control groups) has a proportion of males of around 49%, it seems strange that the proportions in each individual groups was so different. One factor that could have influenced this result is the lack of responses our experiment received.

The control group is about 20% smaller than our treatment group, given that we did not receive answers from that fraction of subjects. A way to minimize this randomization error would have been to use blocking randomization, meaning that we ensured that our random samples had the same proportion of people as they are in the universe. However, although we did think about applying this method during our planning, we saw that it would have required us to run a preliminary survey to learn about the true proportions in our universe. This, we believed, would have been too demanding of the people in our program given that they also had to respond to multiple other teams' surveys and questionnaires. In future experiments, we would allot more time to planning and making sure the groups in the experiment are statistically similar. However, given that our groups do share some significant similarities, we still believe it is valid to run analysis on the treatment effects.

	Treatment	Control	P-value
Under 20	0.0384615	0.1052632	0.515617
Between 20-30	0.9230769	0.8947368	1.000000
Over 30	0.0384615	0.0000000	0.000000

	Treatment	Control	P-value
Male	0.6153846	0.3157895	0.0022497
Female	0.3461538	0.6315789	0.0037487
NonBinary	0.0384615	0.0000000	0.0000000

	Treatment	Control	P-value
Under 3.0	0.0384615	0.0000000	0.0000000
Between 3.0-3.5	0.1923077	0.5263158	0.0006283
Over 3.5	0.7692308	0.4736842	0.0027657

	Treatment	Control	P-value
No Experience	0.3461538	0.3684211	1
Under 2 years	0.4230769	0.4210526	1
Between 2-5	0.1923077	0.2105263	1
Between 6-9+	0.0384615	0.0000000	0

Regression This experiment led to some very interesting findings. Our team ran regressions on two target variables, Score and Completion time of Cognitive test (duration), and added some control variables to analyze the treatment effect resulting from our experiment. As evidenced by the regressions below, analyzing the results of the experiment was difficult given the small sample of responses we received. This is likely the cause of a majority of our values having no statistical significance and a high degree of uncertainty. Regardless, the experiment did provide some insight on our initial question and hypothesis, as seen in the analysis below.

Regression of Score on treatment

```
score_reg <- feols(score~treatment, data=total, se='white')
score_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat","P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	4.526	0.268	16.903	0.000
treatment	-0.603	0.424	-1.422	0.162

```
score_age_reg <- feols(score~treatment + age, data=total, se='white')
score_age_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	2.583	0.713	3.620	0.001
treatment	-0.748	0.391	-1.914	0.063
age>30	2.165	0.749	2.891	0.006
age20-30	2.172	0.723	3.004	0.005

```
score_gender_reg <- feols(score~treatment + gender, data=total, se='white')
score_gender_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	4.481	0.292	15.335	0.000
treatment	-0.678	0.519	-1.305	0.199
genderMale	0.057	0.533	0.108	0.915
genderNon-binary / third gender	2.197	0.508	4.325	0.000
genderPrefer not to say	0.519	0.292	1.777	0.083

```
score_gpa_reg <- feols(score~treatment + GPA, data=total, se='white')
score_gpa_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	3.606	0.424	8.510	0.000
treatment	-0.606	0.424	-1.430	0.160
GPA>3.50	0.984	0.382	2.578	0.014
GPA3.00 - 3.50	0.863	0.412	2.092	0.043

```
score_exp_reg <- feols(score~treatment + work_experience, data=total, se='white')
score_exp_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	4.560	0.429	10.633	0.000
treatment	-0.603	0.434	-1.391	0.172
work_experience0 years	-0.470	0.529	-0.889	0.379
work_experience2 - 5 years	0.664	0.544	1.221	0.229

Predictor	Coefficient	SE	T-Stat	P-Value
work_experience6 - 9 years	0.043	0.409	0.106	0.916

From these regression results it becomes evident that our treatment did not have a statistically significant effect on the scores for people in the treatment group. The only statistically significant values from the regression come from the variables age and GPA. Although some gender variables, non-binary and not disclosed, have what looks like a significant effect on score, it is important to note that in our experiment there was only one instance of either gender, and thus we believe that these coefficients are heavily biased. It is interesting to note that although the effect of the treatment was never significant, it is encouraging to see that the effect of the treatment on scores was negative. This gives us an indication that in a new experiment with a much larger sample size we could potentially find that people do in fact perform worse on the first try if they are aware that they will have multiple opportunities to complete a task.

Regression of completion time on treatment

Our team also sought to analyze the differences our treatment caused on the time it took respondents to finish the cognitive test. Below are the results of the regressions on duration:

```
duration_reg <- feols(duration~treatment, data=total, se='white')
duration_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                    "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	310.474	33.163	9.362	0.000
treatment	4.642	48.536	0.096	0.924

```
duration_age_reg <- feols(duration~treatment + age, data=total, se='white')
duration_age_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient",
               "SE", "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	244.131	97.794	2.496	0.017
treatment	-7.393	49.258	-0.150	0.881
age>30	258.262	100.750	2.563	0.014
age20-30	74.148	99.127	0.748	0.459

```
duration_gender_reg <- feols(duration~treatment + gender, data=total, se='white')
duration_gender_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient",
               "SE", "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	319.023	39.133	8.152	0.000
treatment	-3.832	48.228	-0.079	0.937
genderMale	-2.736	49.398	-0.055	0.956
genderNon-binary / third gender	41.809	35.056	1.193	0.240
genderPrefer not to say	-146.023	39.133	-3.731	0.001

```
duration_gpa_reg <- feols(duration~treatment + GPA, data=total, se='white')
duration_gpa_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient",
              "SE", "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	312.840	45.249	6.914	0.000
treatment	7.160	45.249	0.158	0.875
GPA>3.50	-6.743	41.953	-0.161	0.873
GPA3.00 - 3.50	1.573	42.562	0.037	0.971

```
duration_exp_reg <- feols(duration~treatment + work_experience, data=total, se='white')
duration_exp_reg %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient",
              "SE", "T-Stat", "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	334.571	51.636	6.479	0.000
treatment	-3.350	50.426	-0.066	0.947
work_experience0 years	-40.874	58.760	-0.696	0.491
work_experience2 - 5 years	-42.932	64.313	-0.668	0.508
work_experience6 - 9 years	163.779	56.781	2.884	0.006

From this result we can observe that our treatment had no statistically significant effect on the completion time it took respondents to finish the cognitive test. The coefficient for the treatment in these regressions makes it more difficult to analyze, as for some controls the effect is positive and in others it is the opposite. This is most likely because the treatment has such a negligible effect on the time it took respondents to finish, and other control variables, such as a subject's age or work experience had a bigger effect on the duration of the cognitive test.

Limitations

First, from the test software's online dashboards, we found that some people didn't finish the test after they started. Maybe a part of these people found our IQ quiz questions uncomfortable or bothersome and decided to leave. So,a potential of huge selection bias of only people who enjoy brain teasers and are good at test taking are participating. Second, a sample of college students from all majors in the US would be a more representative sample. However, we only got test results from a part of MSBA students in BU. So, the experiment results will be less applicable to a broader usage in the future.

Another limitation is that the experiment doesn't create a true test environment with supervision. There might be an excludability violation when some test takers google for answers halfway through or taking the test with someone else, even with other test takers.

We only set up tests with less than ten questions due to resource constraints and expected participants' impatience for taking them without money. The test scores contain a lot of noise, much more than they should compared to some standardized tests like SAT or GRE.

Conclusion

This experiment has produced some very interesting findings. After analyzing the effect of our treatment on the duration of the test and the score obtained, we discovered that people in the MSBA program do

not perform differently when they are aware of having multiple chances to complete a task. This finding was surprising as, based on previous experiments and intuition, our team believed that we would see a clear negative effect of the treatment on scores. Although the results were not as expected, some other findings that we did not consider came out of this experiment. The most apparent one, even if evident in hindsight, is the relationship between reported GPA and scores. Even if having a high GPA is not causal to having a higher score, it is very interesting to observe and ponder on why these individuals did statistically significantly better on the cognitive test.

Furthermore, although the treatment effects found were not statistically significant, they nevertheless tended to be in line with our initial hypothesis, which makes us confident to think that in an experiment with a larger sample size we would likely find a significant effect that would support our hypothesis. There are multiple reasons why our experiment did not produce the results we expected it to. Firstly, and as mentioned before, it effects size would have to be immense to be discoverable within a sample of this size. And most importantly, the issue of failure to treat and the overall lack of responses obtained was the biggest crux of our experiment. It is likely that this reduced number of responses contributed to the lack of statistical significance and the statistical differences between treatment and control groups.

In the future, running an experiment similar to this but with some slight adjustments could lead to findings that are both more generalizable and statistically significant. For one, changing the randomization strategy used would lead to more even treatment and control groups. For this, sending out a preliminary survey to know of the proportion of people in the universe with similar characteristics would be allow for Blocking randomization. Moreover, the cognitive test that would be answered could be better planned and tested to allow for a more evident disparity between treatment and control results, as well as having some other controls that could be used to obtain a more precise estimate of the true treatment effect. With these small changes, the experiment would display more significant findings.

Bibliography

- Archer, Kathy. (2018). Do Multiple Homework Attempts Increase Student Learning? A Quantitative Study. The American Economist. 63. 056943451877479. 10.1177/0569434518774790.
- Aimee J. Luebben. (2010). Giving Students Multiple Attempts to Improve Test Scores Provides a Powerful Learning Opportunity.FacultyFocus.