Impact of Test Timer Visibility on Student Test Performance

Final Report

A/B Testing, Design & Implementation - Fall 2019

Group 4:
Tingting Gu
Naphat Korwanich
Honghua Li
Cindy Zhang
Yuwei Zhu

Table of Contents

Abstract

The purpose of this experiment was to examine the effect of timer visibility during an aptitude test for test-takers with an education degree of high school and above. The aptitude test was designed to have a full score of 10 points and be completed within 6 minutes. During the experiment, we set the time limits of one-third of tests to be 4 minutes to impose extra time pressures for test-takers, one-third to be 6 minutes and one-third to be 8 minutes.

Moreover, the time limits are mere guidance for participants; they can still submit a test after the time limit expires. Within the cohort of tests taken under different time limits, the treatment of having a timer visible on top of the screen is applied to half of the tests, while the remaining half of the tests have no timer visible. A participant had an equal probability ⅙ to access any one of the test versions. A total of 173 observations were collected.

The results of a simple linear regression show that having a timer or not does not have a statistically significant effect on the test takers' time spent other than for the 4-minute test groups. Also, a simple linear regression shows that having a timer or not has no statistically significant relationship with the performance on the test. The regression included an interaction term between the time spent on the test, and having a timer visible also did not produce statistically significant results. Lastly, a cohort analysis was also run between test performers with a score above 9, between 6-8, and below 6. The results again proved that the treatment of having a timer or not is statistically insignificant.

Introduction

The time limit has long been used as a tool in exam design. It adds a layer of challenge to questions, and can better evaluate student's performance when being appropriately set[1]. However, inappropriate time limit setting or even the visibility of timer might pose anxiety to the test takers. Thus, the objective of this study is to measure the effect of a visible running countdown on students' logical reasoning performance. To be specific, we would like to examine if a student performs better or worse in basic aptitude tests with a countdown timer visible.

[1] Bridgeman, Brent, et al. "Testing and Time Limits." *ETS R&D Connections*, ETS, 2004, www.ets.org/Media/Research/pdf/RD_Connections1.pdf.

Experimentation Design

Unit of Analysis

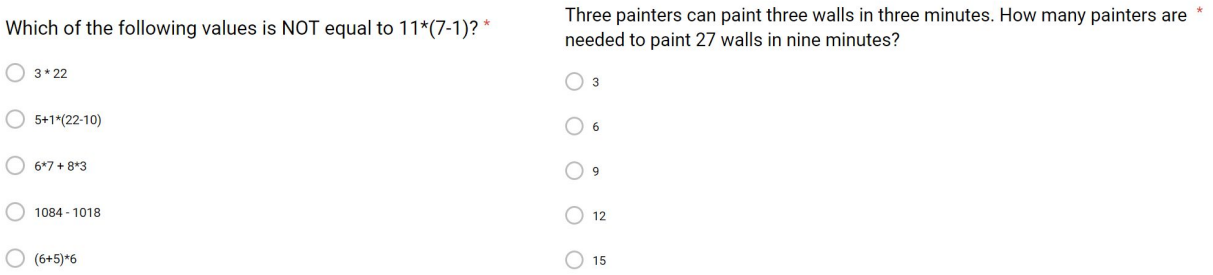
The experiment has been conducted at the individual quiz taker level. It's open to anyone with an education level of high school and above.

Test Question Design

The quiz questions are designed to be concise and avoid linguistic ambiguity. The test topic is taken mainly from the Math test published by the PA government for 8th Graders in 2018[2] as well as logic questions from the Wonderlic Test, a professional IQ test developer[3]. The tests include ten questions and have a total score of ten points.

However, since most questions require a basic understanding of the English language, we also ask if English is the respondent's native language.

The following is an example of 2 out of 10 questions in this quiz:



Treatment & Randomization

The test is designed to be completed within 6 minutes, based on the average time that each team member took. To evaluate the causal effect of the time limit and timer visibility on test performance, we set different versions of the tests by varying the time limits among 4, 6, and 8 minutes. It is worth noting that we chose 4, 6, 8 minutes as intervals to approximate three states of test-taking. Given the average time spent on finishing the test rounds to 4 minutes, we intend

---

[2] "The Pennsylvania System of School Assessment Mathematics Item and Scoring Sampler." Edited by Pennsylvania PSSA, *PSSA Grade 8 Mathematics Item Sampler 2019-2020*, Pennsylvania Department of Education Bureau, 2019, www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Item%20and%20Scoring%20Samples/2019%20PSSA%20ISS%20Math%20Grade%208.pdf.

[3] Wonderlic Test. "Full Quiz." *Full Quiz | Sample Wonderlic Test*, 2019, www.samplewonderlictest.com/quiz/full-quiz.

to use four minutes to mimic the **rushing test-taking state**, 6 minutes to **just in time**, and 8 minutes to **more than enough time**.

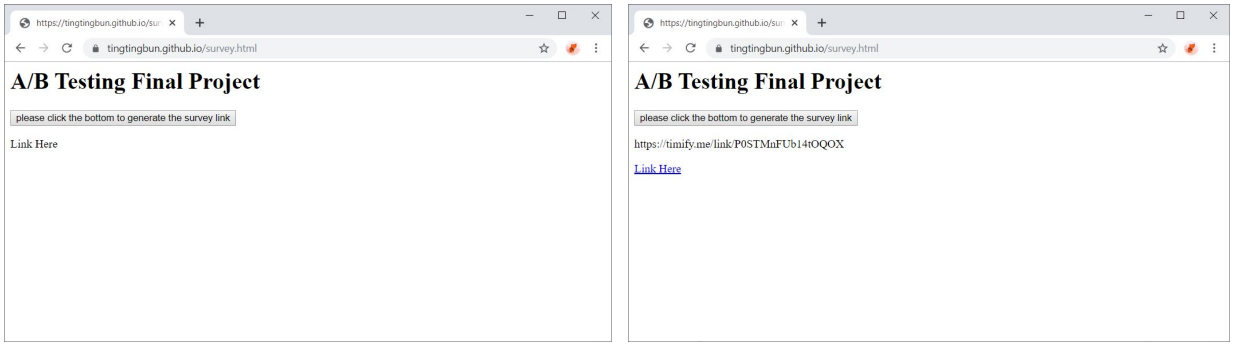
There are six versions of tests: 4-minute tests without a timer, 4-minute tests with a timer; 6-minute tests without a timer, 6-minute tests with a timer; 8-minute tests without a timer, 8-minute tests with a timer. The treatment for each version of the tests is to have a timer visible on the test interface. To randomize the subject into treatment and control groups of different time limits, we used a JavaScript webpage is used to randomly assign test taker to one of the six versions of the quizzes. It will randomly return one link from a list of pre-generated links. The randomization code used is as below:

```
<script>

    function shuffleArray(d) {
     for (var c = d.length - 1; c > 0; c--) {
     var b = Math.floor(Math.random() * (c + 1));
     var a = d[c];
     d[c] = d[b];
     d[b] = a;
            }
      return d
     };
```

The aforementioned pre-generated links were generated using Timify platform[4], a plug-in for Google Forms that used in managing timer capability and generating unique links to the test, as shown in Figure 3.

Figure 1: *Screenshots of the link generation page, and example of when a link gets generated.*



[4] Timify Platform: www.timify.com.

Figure 2: *Screenshots of the test's landing page and the test interface. In this example, the test has a time limit of 6 minutes with the timer visible at the top of the page.*

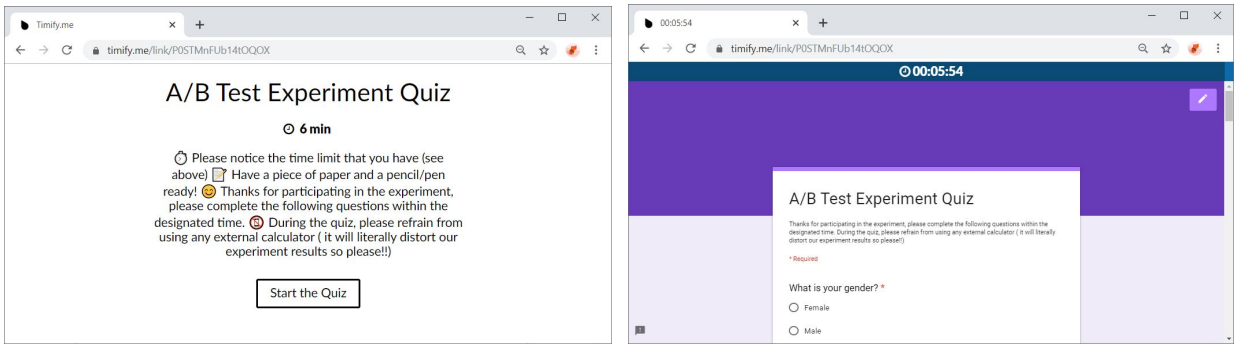
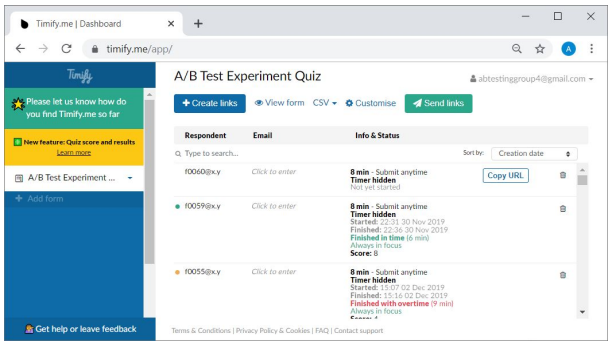


Figure 3: *Screenshot of Timify platform*



Additional Data Collected

In addition to the ten test questions, we also asked for demographic data from test-takers, including gender, age, education, and English proficiency. Detailed analyses of these attributes are described in the next section, "Data Analysis & Modeling."

Platform

There are two key parts of the test-taking platforms:

- Google Forms where the tests are hosted: a few demographic questions and the 10-question aptitude test including mathematics and logic questions;
- Timify: a plug-in that allows the timer capability and containing the Google Forms quiz inside, with the premium feature enabled for the ability to hide or show the timer.

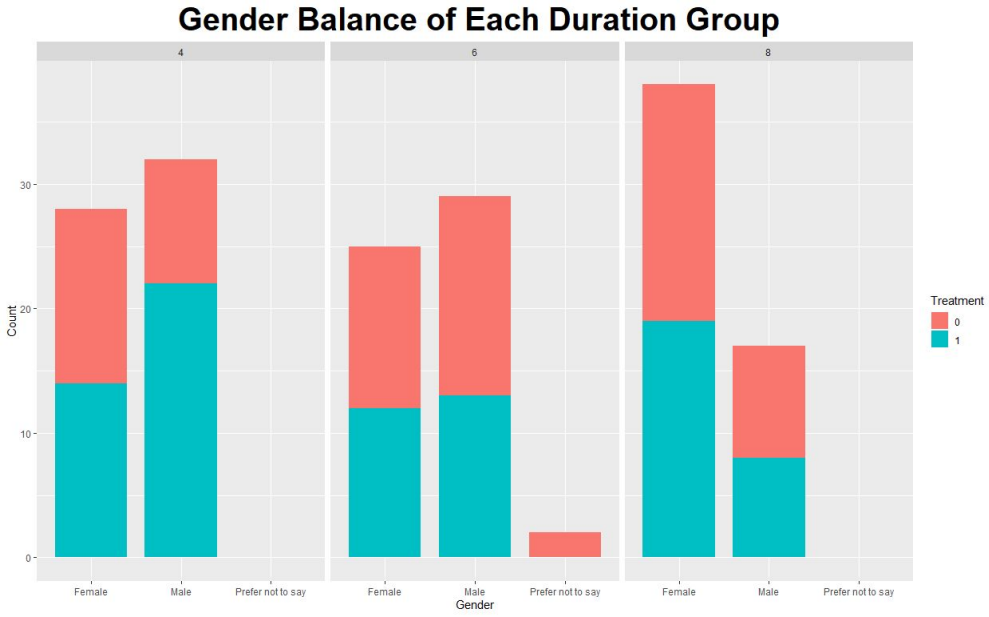
Data Analysis & Modeling

Data Distribution & Balance Check

To check if we have randomized assignments of test-takers between treatment and control groups, a t-test (or Chi-square test of the categorical test) was performed on control variables: gender, age, education, and if an English native speaker.

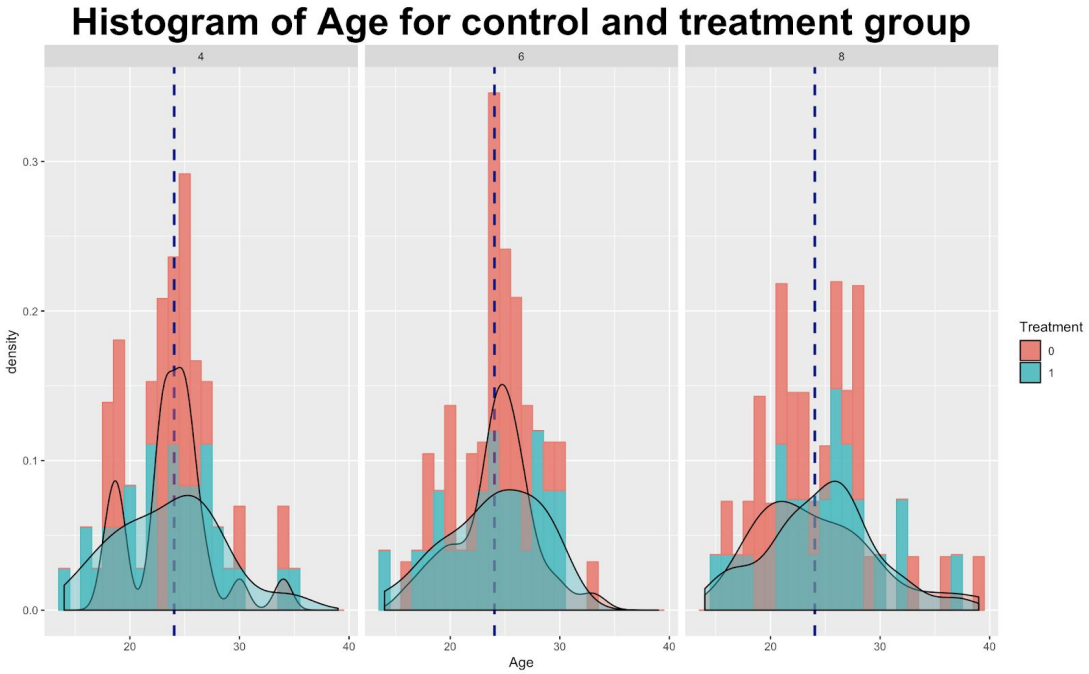
Besides, we performed t-tests for total time spent on the tests and total scores participants achieved. On a statistically significant level, the time spent on the tests is longer for the control group (without a timer visible) than the treatment group (with timer visible).

Gender



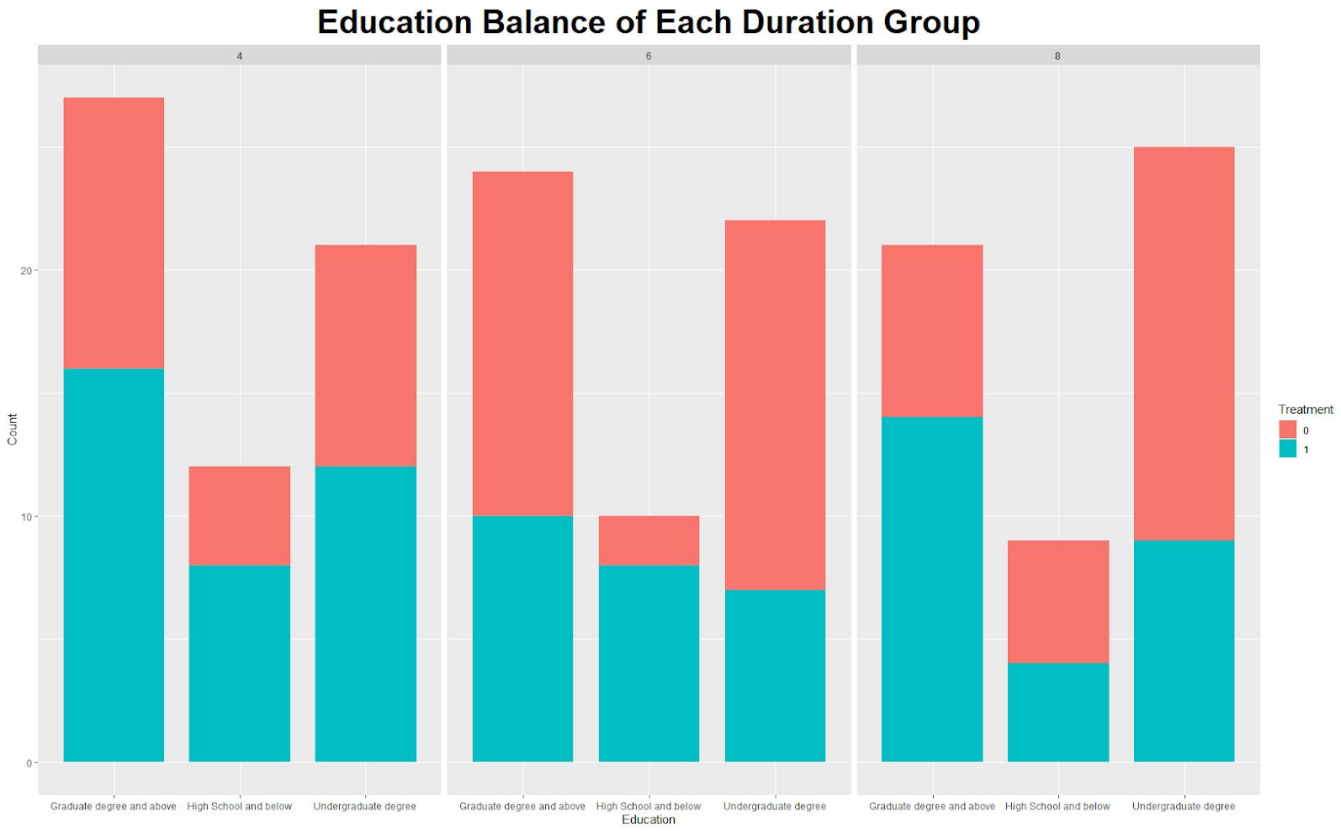
	# of Female		# of Male		# of Prefer not to say		P-value (Chi-Square test)
	Treatment	**Control**	**Treatment**	**Control**	**Treatment**	**Control**	
Overall	45	46	43	35	0	2	0.2609
4-minute group	14	14	22	10	0	0	0.2244
6-minute group	12	13	13	16	0	2	0.4216
8-minute group	19	19	8	9	0	0	1

Age



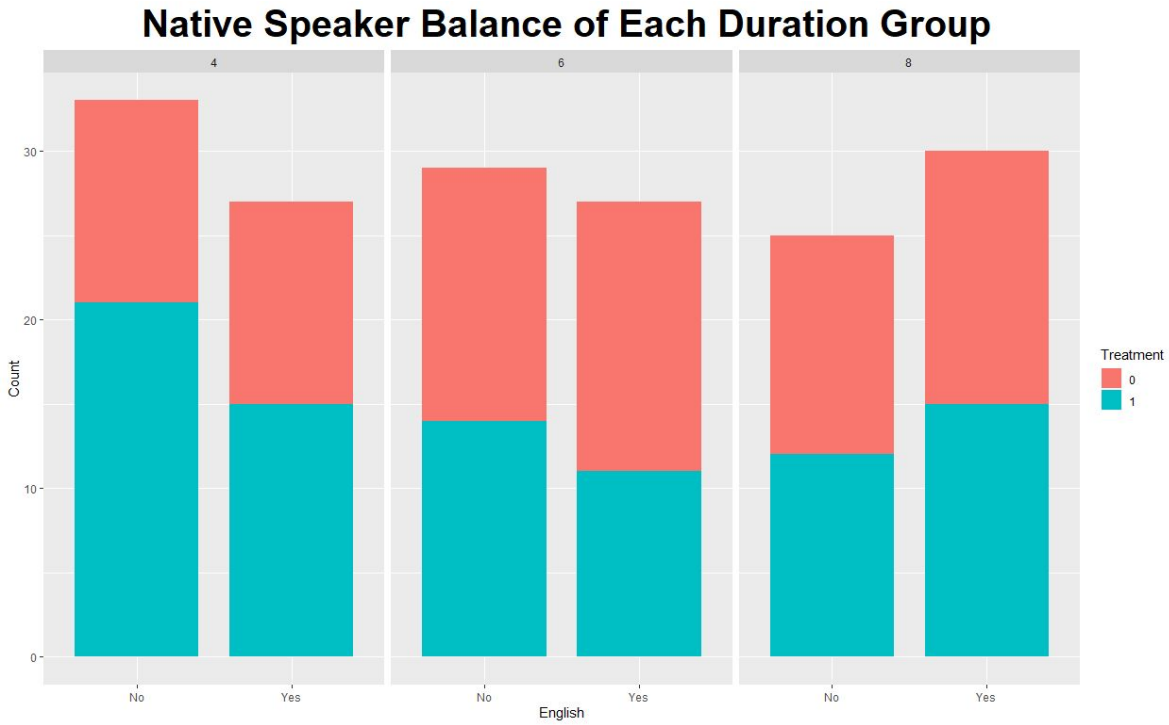
	Treatment Group		Control Group		P-value (Chi-Square test)
	Mean	**Median**	**Mean**	**Median**	
Overall	23.98864	24	24.08	24	0.8904
4-minute group	23.52	24	23.75	24	0.8415
6-minute group	23.96	24	24.16	24	0.3273
8-minute group	24.62	25	24.28	23	0.709

Education



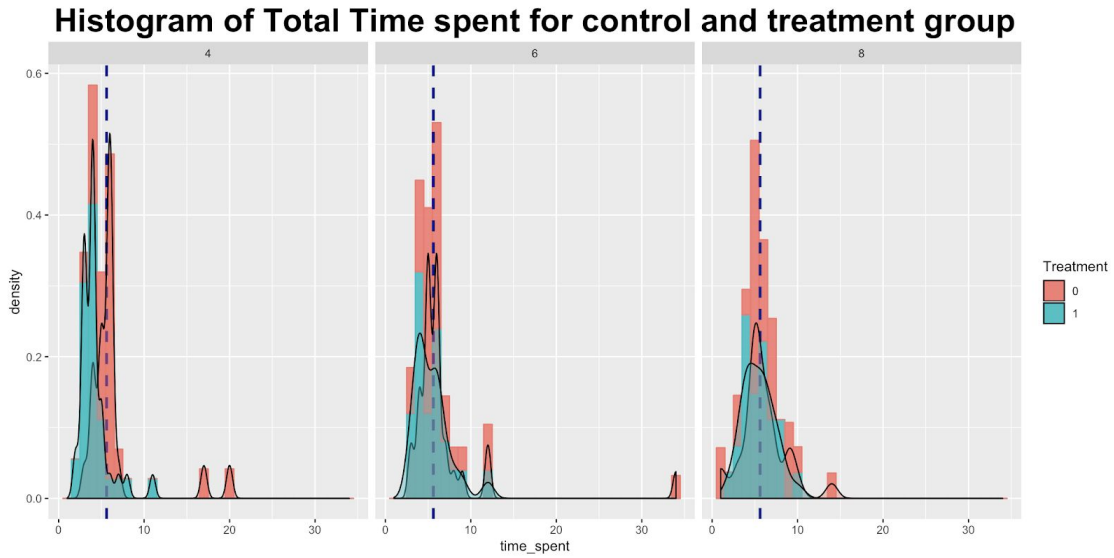
	# High school and below		# of Undergraduate Degree		# of Graduate and above		P-value (Chi-Square test)
	Treatment	**Control**	**Treatment**	**Control**	**Treatment**	**Control**	
Overall	20	11	28	40	40	32	0.06464
4-minute group	8	4	12	9	16	11	0.8608
6-minute group	8	2	7	15	10	14	0.03672
8-minute group	4	5	9	16	14	7	0.1115

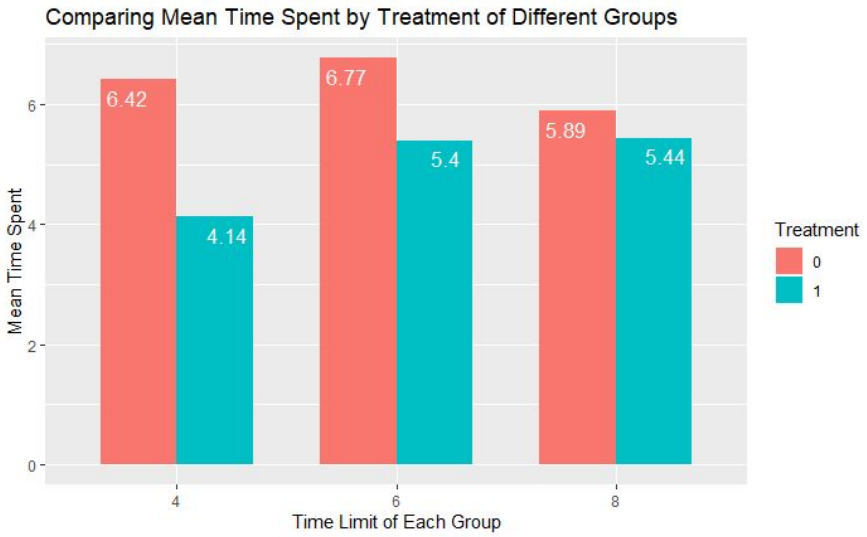
English Proficiency

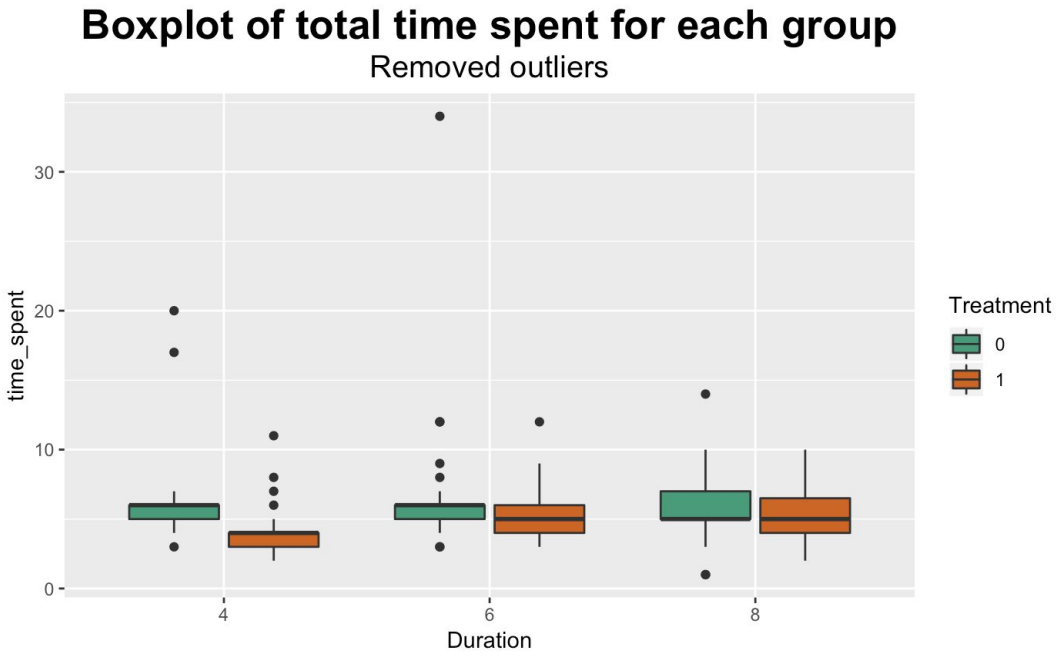


	# of Native Speaker		# of Not Native		P-value (Chi-Square test)
	Treatment	**Control**	**Treatment**	**Control**	
Overall	41	43	47	40	0.5969
4-minute group	15	12	21	12	0.7108
6-minute group	11	16	14	15	0.7659
8-minute group	15	15	12	13	1

Total Time Spent

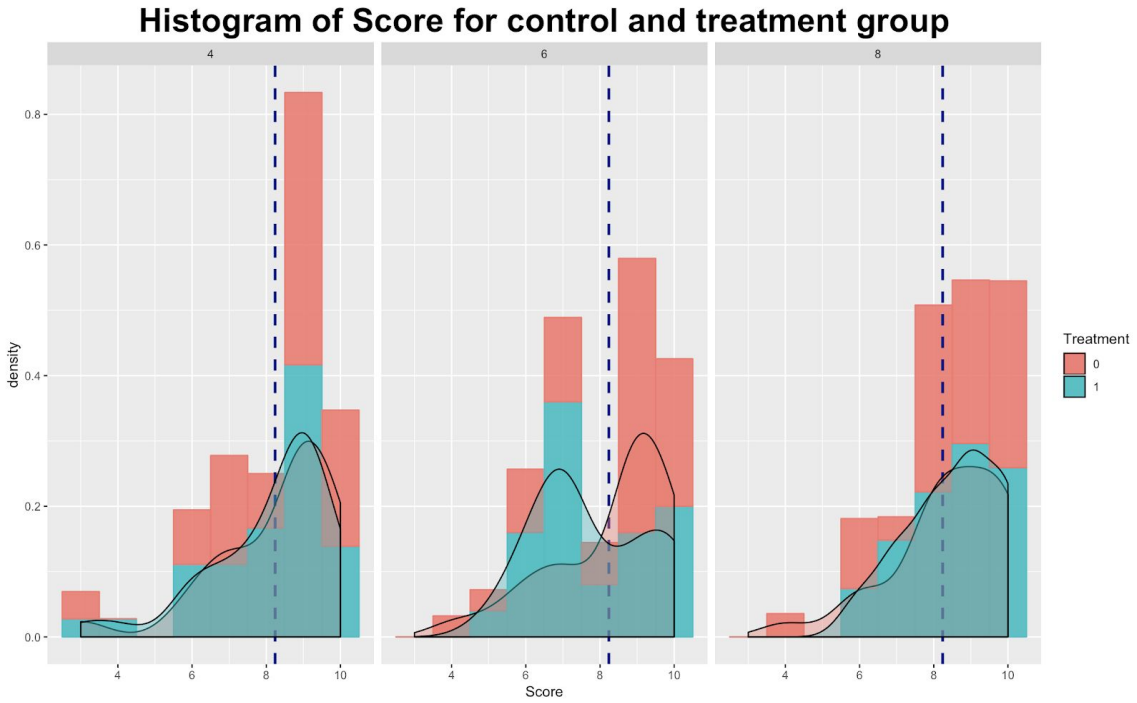


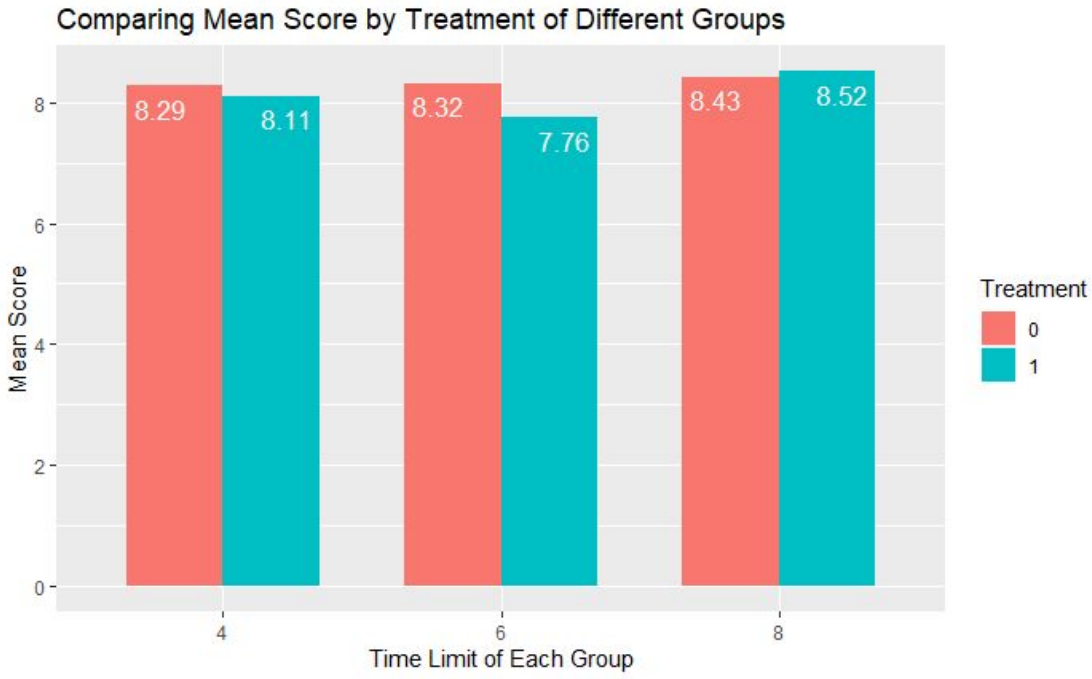


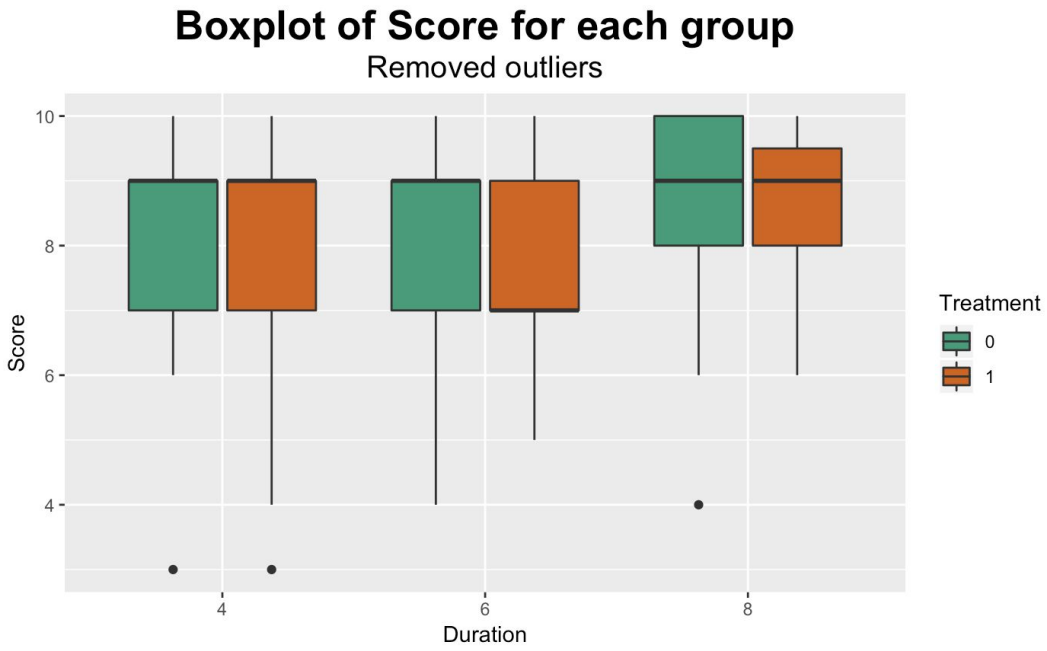


	Treatment Group		Control Group		P-value (Chi-square test)
	Mean	**Median**	**Mean**	**Median**	
Overall	4.897727 mins	4 mins	6.373494 mins	6 mins	0.004183
4-minute group	4.138889 mins	4 mins	6.416667 min	6 mins	0.01099
6-minute group	5.400000 mins	5 mins	6.774194 min	6 mins	0.6302
8-minute group	5.444444 mins	5 mins	5.892857 mins	5 mins	0.1675

Score


Histogram of Score for control and treatment group
4
6
8
density
0.0
0.2
0.4
0.6
0.8
4
6
8
10
Score
Treatment
0
1


Comparing Mean Score by Treatment of Different Groups
Mean Score
0
2
4
6
8
8.29
8.11
8.32
7.76
8.43
8.52
4
6
8
Time Limit of Each Group
Treatment
0
1



	Treatment Group		Control Group		P-value (Chi-Square test))
	Mean	**Median**	**Mean**	**Median**	
Overall	8.136364	9	8.349398	8.5	0.3717
4-minute group	8.111111	9	8.291667	9	0.6833
6-minute group	7.760000	7	8.322581	9	0.4568
8-minute group	8.518519	9	8.428571	9	0.5886

Results

Regression Introduction:
Treatment: (Test with timer) = 1 (Test without timer) = 0
Standard Error: The standard error is clustered on each individual

Baseline Linear Regression of Score on Treatment

Before we perform a more in-depth analysis to understand the treatment effect of having a timer on tests, we would use a baseline model to control for covariates such as gender, education, English proficiency, which bring inherent variation to each experimental subject.

Timer effect on 3 different groups

	Dependent variable:			
	Score			
	Four minute (1)	Six minute (2)	Eight minute (3)	Overall (4)
factor(Treatment)1	-0.283	-0.160	0.386	-0.111
	(0.401)	(0.453)	(0.433)	(0.243)
factor(Gender)Male	0.186	0.331	0.454	0.341
	(0.464)	(0.452)	(0.393)	(0.242)
factor(Gender)Prefer not to say		1.148		1.239**
		(0.953)		(0.559)
factor(Education)High School and below	1.448	-1.181**	0.423	0.122
	(0.881)	(0.528)	(0.618)	(0.400)
factor(Education)Undergraduate degree	0.719	0.406	0.788	0.452
	(0.839)	(0.505)	(0.543)	(0.321)
factor(English)Yes	-0.384	-0.061	0.038	0.106
	(0.796)	(0.438)	(0.468)	(0.291)
Age	0.075	-0.128***	0.025	-0.001
	(0.050)	(0.045)	(0.035)	(0.026)
Constant	6.120***	11.100***	7.028***	7.891***
	(1.239)	(1.273)	(1.139)	(0.695)
Observations	60	57	56	173
R2	0.077	0.172	0.086	0.041
Adjusted R2	-0.028	0.053	-0.025	0.001
Residual Std. Error	1.675 (df = 53)	1.542 (df = 49)	1.468 (df = 49)	1.563 (df = 165)
F Statistic	0.735 (df = 6; 53)	1.450 (df = 7; 49)	0.772 (df = 6; 49)	1.014 (df = 7; 165)
Note:				*p<0.1; **p<0.05; ***p<0.01

The linear regression above shows that the average treatment effect of having a timer across all three time limits is not statistically significant, regardless of the magnitude of the coefficients. In this overview, since we are only concerning the average treatment effect of having a timer (instead of a 4-minute timer, 6-minute timer, 8-minute timer), we subset our population subjects based on the time limit.

Base-line Linear Regression of time_spend on Treatment

Time_spend: total time a subject spent on the test*

*subjects are able to work over the time limit for both control and treatment group

Timer effect on 3 different groups

	Dependent variable:			
	time_spent			
	Four Minute (1)	Six Minute (2)	Eight Minute (3)	Overall (4)
factor(Treatment)1	-2.421***	-18.168	-3.106	-8.011
	(0.839)	(17.706)	(2.895)	(5.947)
factor(Gender)Male	0.003	-23.442	-1.883	-6.784
	(0.665)	(22.437)	(1.575)	(5.783)
factor(Gender)Prefer not to say		-48.963		-13.314
		(48.275)		(12.607)
factor(Education)High School and below	2.358*	2.943	-2.589	-1.051
	(1.363)	(18.218)	(3.028)	(1.967)
factor(Education)Undergraduate degree	2.315*	16.602	-3.285	4.089
	(1.376)	(19.607)	(3.629)	(5.495)
factor(English)Yes	-1.913*	28.939	-1.787	5.435
	(1.032)	(28.812)	(1.279)	(6.822)
Age	0.231	-0.553	-0.185	0.045
	(0.149)	(1.884)	(0.147)	(0.171)
Constant	0.637	34.209	14.169*	11.551*
	(3.582)	(35.948)	(8.183)	(5.945)
Observations	60	57	56	173
R2	0.276	0.071	0.098	0.023
Adjusted R2	0.194	-0.062	-0.012	-0.018
Residual Std. Error	2.662 (df = 53)	82.660 (df = 49)	8.297 (df = 49)	46.757 (df = 165)
F Statistic	3.370*** (df = 6; 53)	0.536 (df = 7; 49)	0.890 (df = 6; 49)	0.563 (df = 7; 165)
Note:				*p<0.1; **p<0.05; ***p<0.01

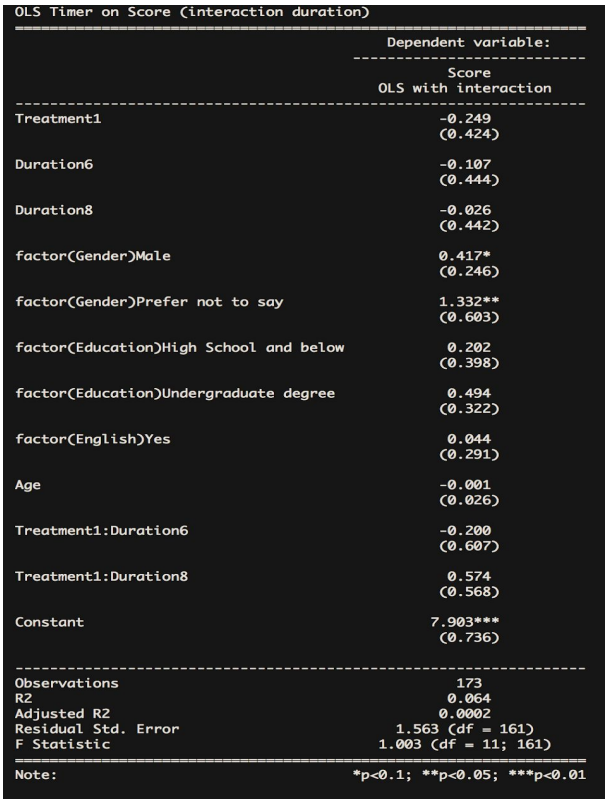
We further regressed time_spend on the treatment across the tests with three different time limits. We found that although the coefficient is not statistically significant for tests with time limit 6 minutes and 8 minutes, having a timer is statistically impactful in terms of reducing the total time the experimental subjects spent on the test for 4 minutes.

The significance in imposing a timer on the 4-minute group indicates that when test takers are rushing, they are more inclined to submit the test quickly. On the contrary, when test takers are given enough time, seeing a timer does not impact how they allocate their time and the total time spent on a test.

Linear Regression of Score on Treatment (Including Duration: Treatment as Interaction Variable)

In the baseline linear regression, we only focused on the treatment effect of having a timer regardless of the time duration. The previous regression showed that there might be a significant relationship between the total time spent and the presence of a timer. Therefore, in this regression, we drill down to the various treatment effect of a timer with different time limits by adding an interaction term between the time limits of the test and the treatment. The time limits are labeled as “Duration” as a categorical variable in the regression below.

The regression we perform is as following:
*ols_interaction <-lm(Score ~ Treatment + Duration + Treatment*Duration + factor(Gender) + factor(Education) + factor(English) + Age, data = MyData)*



OLS Timer on Score (interaction duration)

	Dependent variable:
	Score
	OLS with interaction
Treatment1	-0.249
	(0.424)
Duration6	-0.107
	(0.444)
Duration8	-0.026
	(0.442)
factor(Gender)Male	0.417*
	(0.246)
factor(Gender)Prefer not to say	1.332**
	(0.603)
factor(Education)High School and below	0.202
	(0.398)
factor(Education)Undergraduate degree	0.494
	(0.322)
factor(English)Yes	0.044
	(0.291)
Age	-0.001
	(0.026)
Treatment1:Duration6	-0.200
	(0.607)
Treatment1:Duration8	0.574
	(0.568)
Constant	7.903***
	(0.736)
Observations	173
R2	0.064
Adjusted R2	0.0002
Residual Std. Error	1.563 (df = 161)
F Statistic	1.003 (df = 11; 161)
Note:	*p<0.1; **p<0.05; ***p<0.01

An interaction variable between Treatment and Duration is now included to evaluate the treatment effect for having a timer for X minute test. Similar to the previous regressions, we do not see a statistically significant impact of imposing timer with any specific test duration.

Linear Regression of Score on Treatment (Cohort Analysis)

To increase the granularity of our analysis, we identify three cohorts in our experimental subject. The three cohorts are as follows:
Cohort 1: test-takers with a score below 6,
Cohort 2: test-takers with a score between 6 to 8, and
Cohort 3: test-takers with a score equal to or above 9.

There is a limitation in our cohort design since our treatment swung the test scores; however, we assume that the scores captured some unobserved covariates associated with test-takers. For instance, Cohort 3 may generally perform better in tests than other cohorts.

	Dependent variable:		
	Score		
	OLS Cohort(equal or below 6)	OLS Cohort(equal or below 8)	OLS Cohort(equal or above 9)
	(1)	(2)	(3)
Treatment1	-0.363	0.353	-0.118
	(1.338)	(0.252)	(0.141)
Duration6	0.102	0.031	0.013
	(1.206)	(0.316)	(0.163)
Duration8	-0.111	0.585**	0.153
	(1.185)	(0.268)	(0.187)
Age	0.010	0.008	0.001
	(0.052)	(0.016)	(0.013)
factor(Education)High School and below	1.245***	0.251	0.465**
	(0.388)	(0.231)	(0.184)
factor(Education)Undergraduate degree	0.729	0.274*	0.298**
	(0.627)	(0.163)	(0.135)
factor(English)Yes	0.150	-0.007	-0.158
	(0.493)	(0.158)	(0.123)
Treatment1:Duration6	0.076	-0.478	0.286
	(1.356)	(0.361)	(0.262)
Treatment1:Duration8	1.528	-0.635*	0.129
	(1.644)	(0.336)	(0.245)
Constant	4.674***	6.930***	9.208***
	(1.803)	(0.418)	(0.341)
Observations	26	52	95
R2	0.271	0.297	0.156
Adjusted R2	-0.139	0.146	0.067
Residual Std. Error	1.053 (df = 16)	0.467 (df = 42)	0.474 (df = 85)
F Statistic	0.660 (df = 9; 16)	1.967* (df = 9; 42)	1.747* (df = 9; 85)
Note:			*p<0.1; **p<0.05; ***p<0.01

The results again show an insignificant average treatment effect across three time durations. However, the insignificance could be attributed to our limited sample size. If these coefficients were proven significant with larger sample sizes, we could find that seeing a timer reduces scores for Cohort 3 and Cohort 1 while it improves performance for cohort 2. One interpretation is that subjects under Cohort 3 tend to concern more about test results; the visibility of a timer, therefore, stimulates more stress during tests, resulting in a lower score. For Cohort 2 and 1, which we consider as average test takers and below-average, respectively, the timer may provide positive nudges leading to an improved score. With that being said, these interpretations are limited by both the insignificant causal effect and our cohort design.

Conclusion

The experiment shows that there is no significant relationship between having a timer visible and the results of a participant's performance. A future improvement would be to obtain a larger sample size and perform propensity score matching to estimate the causal effect of having a visible timer on subjects' performance. A more careful cohort analysis determined by participants' prior testing performance could enrich our analyses and improve the quality of our experiment.

Limitations

Due to limited technical capability of Google Forms and Timify platforms, there are two key limitations that we could improve upon in the next iteration of the experimentation.

- **Lack of auto-submission capability** -- Since Timify cannot force the test taker to auto submit the test once the time limit reached, overtime submission was allowed. This could impact the significance of the treatment effect, especially on the 4-min version of the test.
- **Lack of capability to record the time spent on each question** -- There is room to improve the analysis by understanding test-takers' performance on each specific question apart from only looking at total scores.

References

Bridgeman, Brent, et al. “Testing and Time Limits.” *ETS R&D Connections*, ETS, 2004, www.ets.org/Media/Research/pdf/RD_Connections1.pdf.

Wonderlic Test. “Full Quiz.” *Full Quiz | Sample Wonderlic Test*, 2019, www.samplewonderlictest.com/quiz/full-quiz.

“The Pennsylvania System of School Assessment Mathematics Item and Scoring Sampler.” Edited by Pennsylvania PSSA, *PSSA Grade 8 Mathematics Item Sampler 2019-2020*, Pennsylvania Department of Education Bureau, 2019, www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Item%20and%20Scoring%20Samples/2019%20PSSA%20ISS%20Math%20Grade%208.pdf.

Timify Platform: www.timify.com.