Carnegie Mellon University

Does a Visible Timer affect Performance?

A/B Testing, Design & Implementation - Fall 2019

Tingting Gu, Naphat Korwanich, Honghua Li, Cindy Zhang, Yuwei Zhu

Agenda

- Motivation
- Experiment Design
- Data Analyses
- Modeling and Results
- Limitations
- References

Carnegie Mellon University

Motivation of Experiment

Previous Research Focused on Effects of Having a Time Limit on Students' Performance

Extending the SAT tests by 50% longer did not significantly increase the students' performance

---

What would the impact of having a timer visible during a test that measures subjects' basic aptitude abilities?

Hypothesis: with a timer being visible, test takers would be stressed that they would perform worse than test takers without a timer.

Carnegie Mellon University

Carnegie Mellon University

Experiment Design and Execution

Experiment Design

Unit of Analysis

Individual Test Taker

Test Questions

10 questions in total, to be completed in 6 min; adapted from PA High School Math Assessments and Professional Logic Tests; demographic information: age, gender, education, English proficiency

Treatment

6 versions of tests:

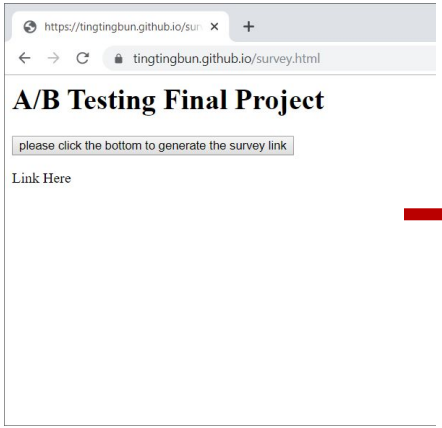
Time Limit	Control/Treatment
4 min – "Speed"	Timer Invisible / Visible
6 min – Enough	
8 min – More than Enough	

Carnegie Mellon University

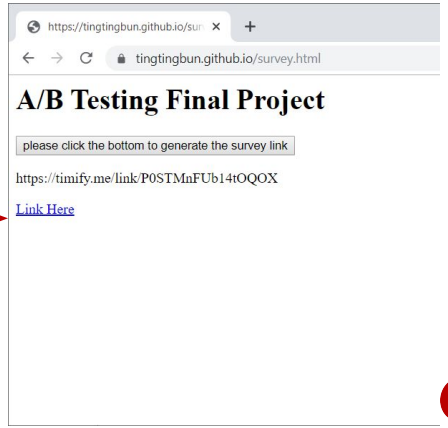
Experiment Design

Randomization

A JavaScript webpage is used to randomly assign test taker to one of the 6 versions of the quizzes

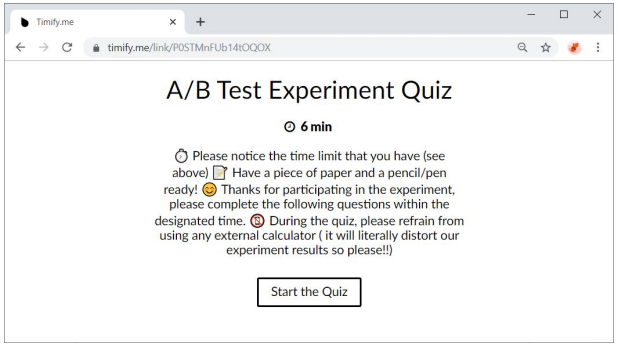


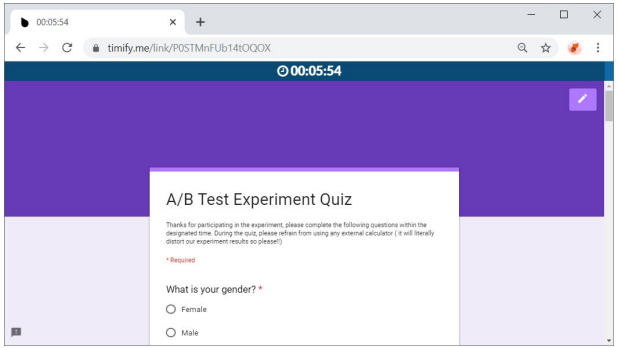

```
<script>

    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
     };
```



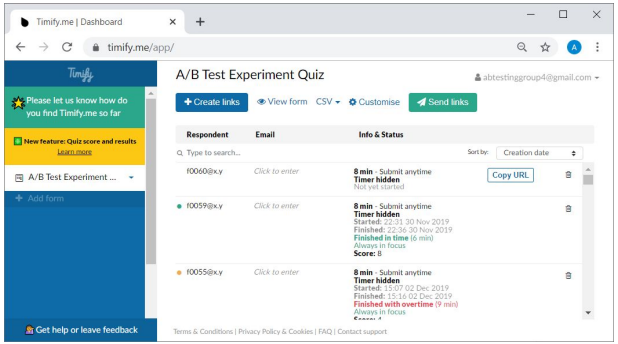
Carnegie Mellon University

Test Execution and Results Collection





Tests are hosted using Google Form and Timify plugin to show a timer when needed



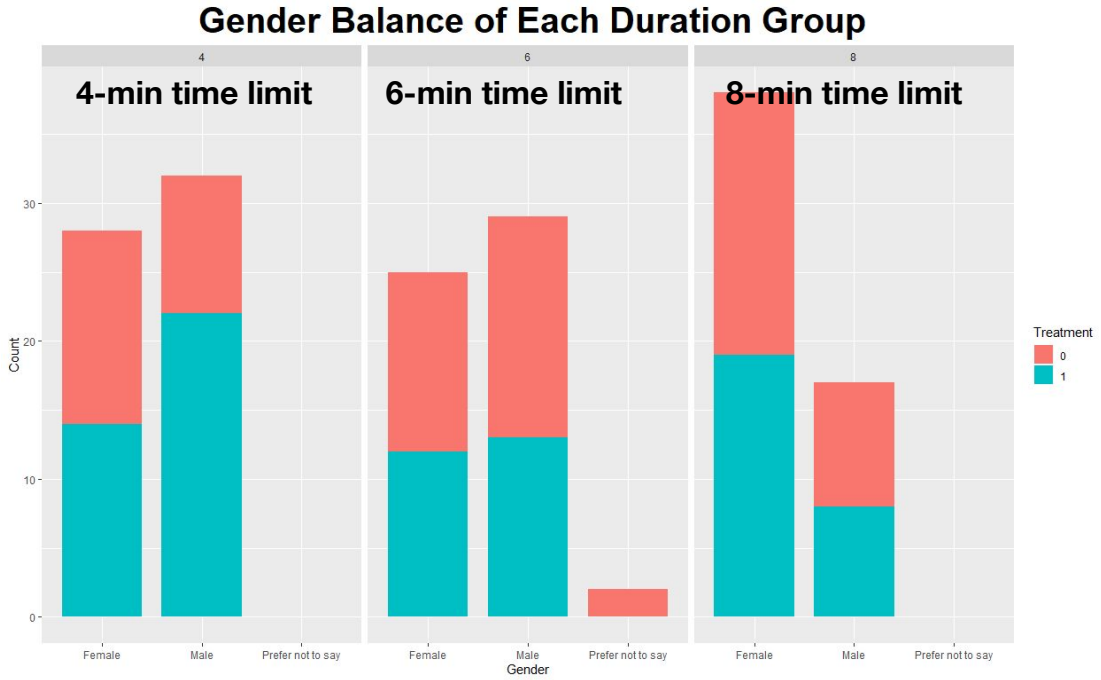
Results are collected and exported on the Timify platform

Carnegie Mellon University

Carnegie Mellon University

Data Analyses

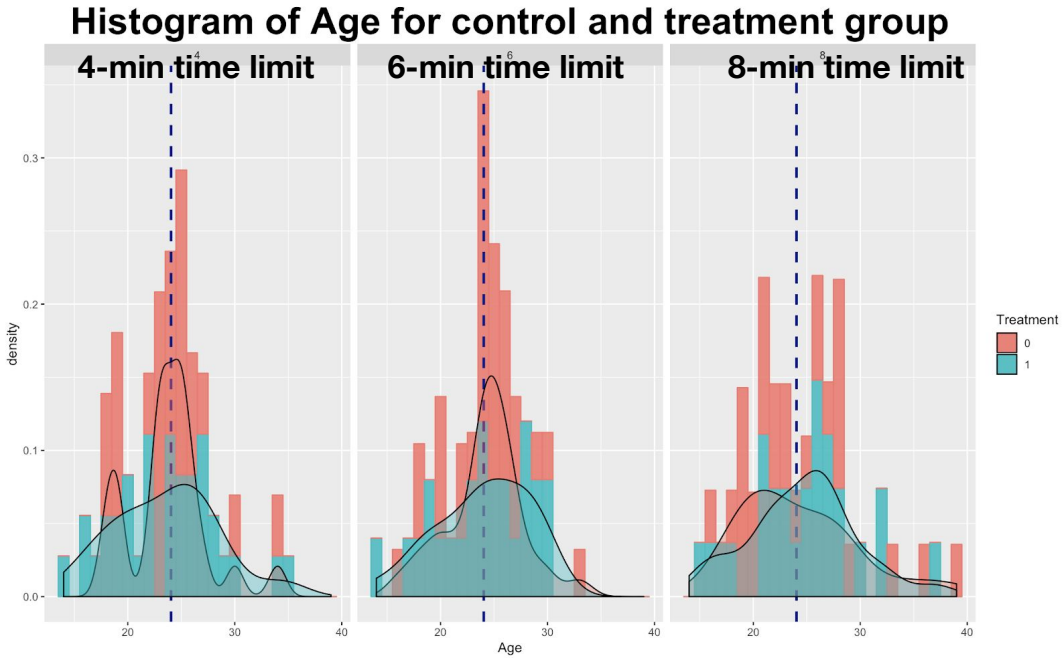
There is no gender imbalance among treatment/ control groups



Groups	P-value
Overall	0.2609
4-minute group	0.2244
6-minute group	0.4216
8-minute group	1

Carnegie Mellon University

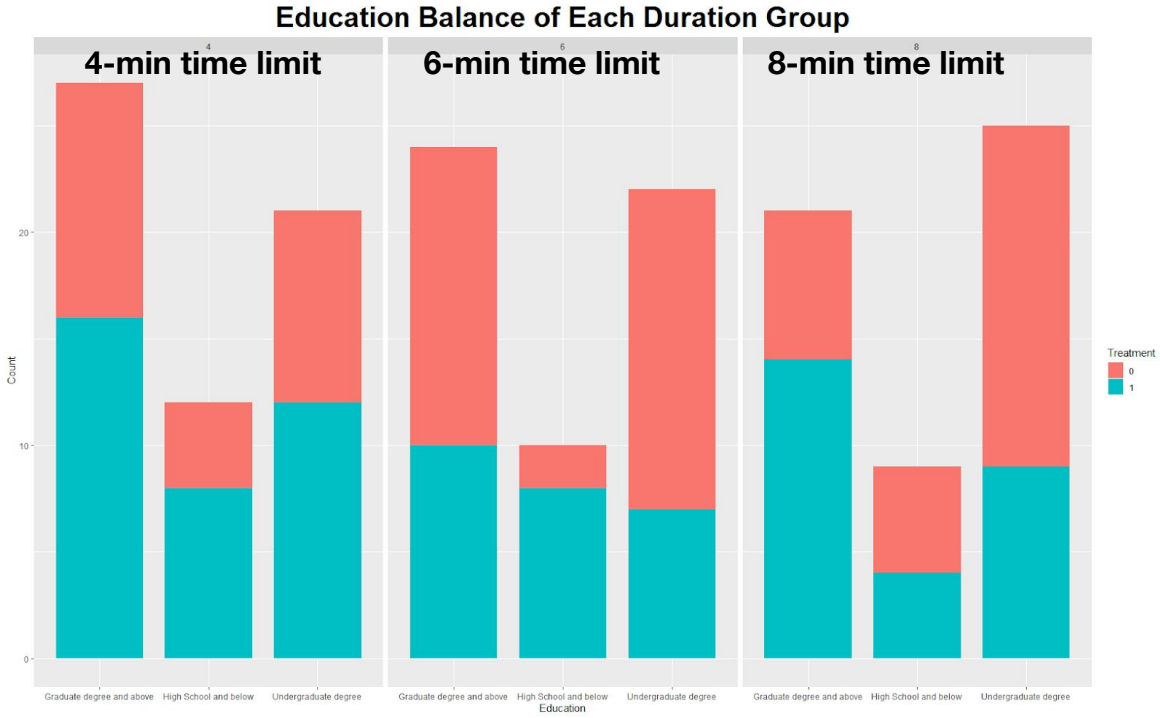
There is no age imbalance among treatment/control groups



Groups	P-value
Overall	0.8904
4-minute group	0.8415
6-minute group	0.3273
8-minute group	0.709

Carnegie Mellon University

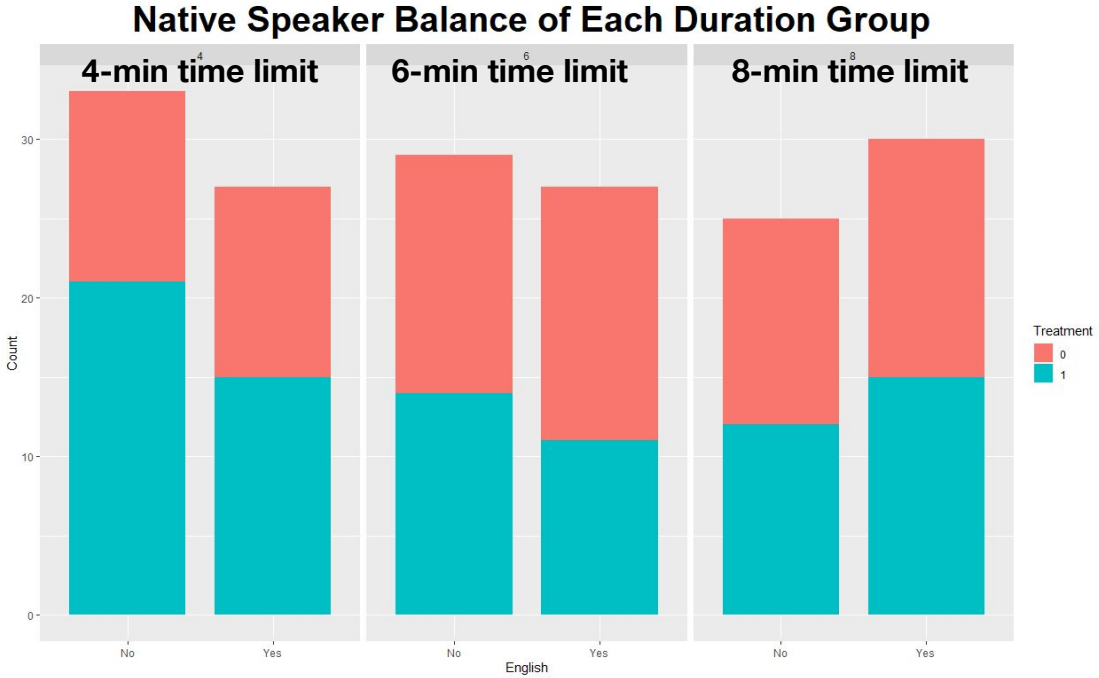
There is education imbalance among 6-min groups



Groups	P-value
Overall	0.06464
4-minute group	0.8608
6-minute group	0.03672
8-minute group	0.1115

Carnegie Mellon University

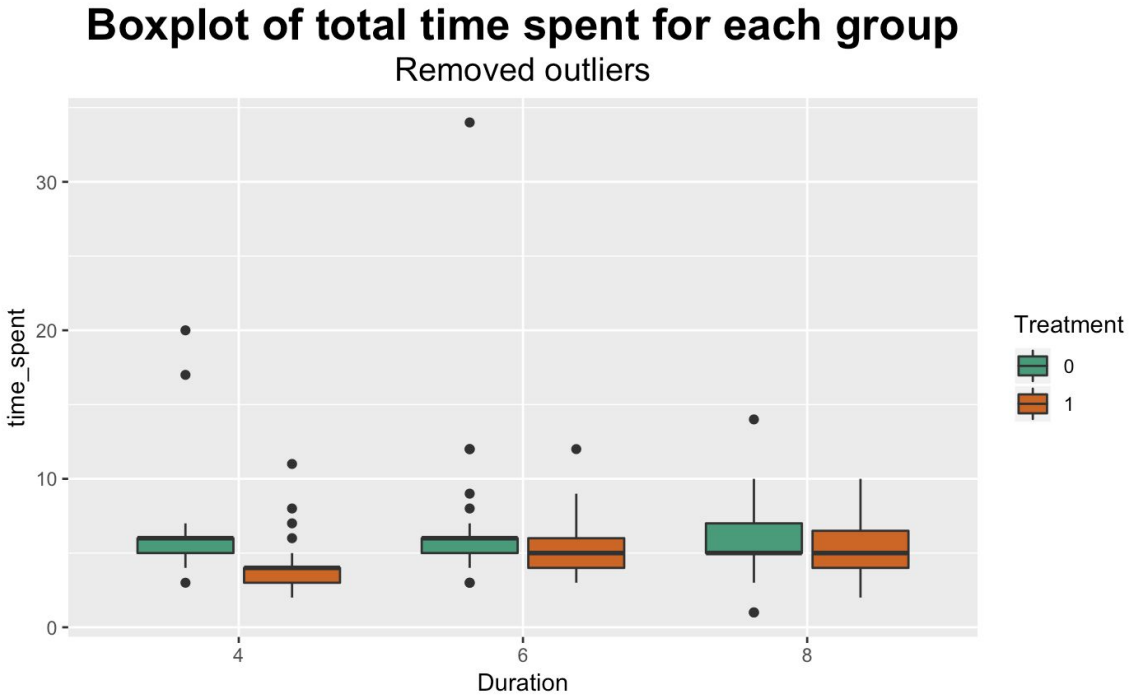
There is no English proficiency imbalance among treatment/control groups



Groups	P-value
Overall	0.5969
4-minute group	0.7108
6-minute group	0.7659
8-minute group	1

Carnegie Mellon University

Total time spent is affected by visibility of a timer

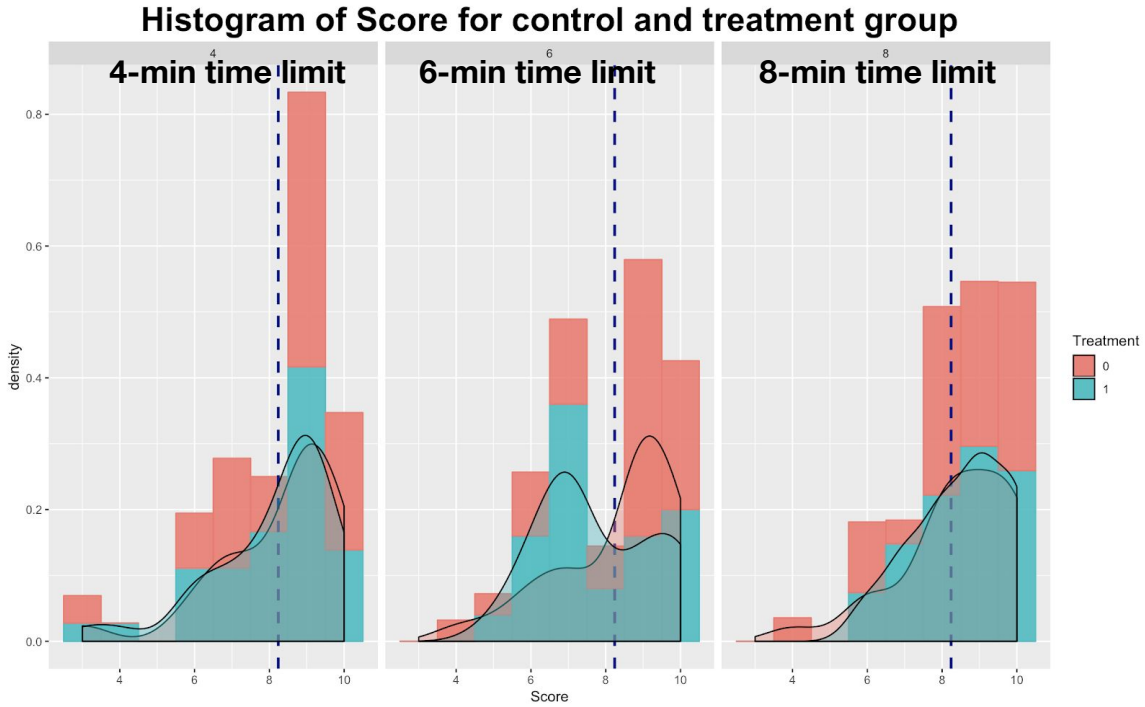


Groups	P-value
Overall	0.004183
4-minute group	0.01099
6-minute group	0.6302
8-minute group	0.1675

Test takers under 4-minute visible timer were under the rush and most inclined to submit tests quickly

Carnegie Mellon University

The score differences among treatment/control groups are insignificant



Groups	P-value
Overall	0.3717
4-minute group	0.6833
6-minute group	0.4568
8-minute group	0.5886

Carnegie Mellon University

Carnegie Mellon University

Modeling and Results

Modeling Overview

	Regression	Results
1	Linear regression of score over treatment and demographic variables across time limits	Insignificant coefficients of treatment
2	Linear regression of time spent over treatment and demographic variables	Insignificant coefficients of treatment other than the 4-min group
3	Linear regression of score over treatment, interaction between time limit and treatment and other demographic variables	Insignificant coefficients of treatment
4	Cohort Analysis (by score range*)	Insignificant coefficients of treatment

1 There is no significant relationship between timer visibility and test performance

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

Average treatment effect of having a timer across 3 time limit groups is not statistically significant, regardless of the magnitude of the coefficients

Carnegie Mellon University

2 There is no significant relationship between timer visibility and total time spent other than the 4 minute 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	-3.186	-8.011
	(0.839)	(17.706)	(2.895)	(5.947)
factor(Gender)Male	0.003	-23.442	-1.803	-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	2.350*	2.943	-2.589	-1.051
	(1.363)	(10.218)	(3.028)	(1.967)
factor(Education)Undergraduate degree	2.315*	16.602	-3.285	4.089
	(1.376)	(19.697)	(3.629)	(5.495)
factor(English)Yes	-1.913*	20.939	-1.707	5.435
	(1.032)	(20.812)	(1.279)	(6.022)
Age	0.231	-0.553	-0.105	0.045
	(0.149)	(1.084)	(0.147)	(0.171)
Constant	0.637	34.209	14.169*	11.551*
	(3.502)	(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

People under the rushing state are more inclined to submit tests quickly

Carnegie Mellon University

3 There is no significant relationship between timer visibility and test performance with interaction term

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

The timers with specific durations are also statistically insignificant in its impact on test performances.

Carnegie Mellon University

4 There is no significant relationship between timer visibility and test performance in cohort analysis

	Dependent variable:		
	Score		
	OLS Cohort(equal or below 6) (1)	OLS Cohort(equal or below 8) (2)	OLS Cohort(equal or above 9) (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

Cohort design:

Score ≤ 6

6 < score ≤ 8

score ≥ 9

Coefficients yield interesting interpretations but are limited by statistical significance and cohort design

Carnegie Mellon University

Carnegie Mellon University

Limitations

Technical Limitations

Limitations	Implications
Lack of auto-submission capability	Since Timify cannot force the test taker to auto submit the test once the time limit reached, overtime submissions were 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.

Carnegie Mellon University

Carnegie Mellon University

References

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.

Carnegie Mellon University