Project Report

Team #2

Ajay Valecha, Arun Sharma, Sanjay Renduchintala, Jianyu Wang, Kaiqi Zhong, Prasun Shrestha

A/B Testing

Fall 2020

The Effect of On-screen Timer on Reading Comprehension

Executive Summary:

Most universities rely on standardized scores such as GRE and GMAT to assess an individual's reading comprehension skills. In a stressful environment, such as standardized tests, a good user interface (UI) design can play an important role in creating a less stressful exam environment to help the students focus on the task at hand. In our experiment, we look at how a visible on-screen timer affects reading comprehension in the context of a timed assessment. We set up a randomized controlled trial and measure the impact of the visible timer on participants' performance. Our findings suggest that while the visible timer does not have a statistically significant effect on students' final scores, it does impact their reading time, survey completion rate, and question completion time.

1. Introduction

Students require a high level of reading comprehension to excel in higher education. For this reason, most universities require standardized tests such as GRE or GMAT to gauge students' reading comprehension levels. Standardized assessment tests are not always a perfect reflection of students' capabilities as the test scores can be affected by multiple factors unrelated to the student's capabilities such as time pressure and anxiety. A good user interface (UI) design can play an important role in creating a less stressful exam environment to help the students focus on the task at hand. The study of the effects of UI design elements on student performance on the assessments has been of interest to experimental psychologists and cognitive scientists to improve the quality of assessments.[1]

In this paper, we are particularly interested in understanding the effect of having a visible timer on reading compression in the context of timed assessments. Our experiment is designed to test if there is a causal relationship between the presence of a visible timer on the screen and student performance on a digital assessment task. Conflicting anecdotal references such as students feeling a sense of control or the opposite — being distracted or hurried by the timer — have motivated us to set up a randomized experiment to uncover the true relationship between an on-screen timer and student performance.

For this experiment, we used a sample GRE reading comprehension passage and randomized our subjects into two assessment environments — test and control. Both the groups performed the assessment task in environments that are identical with an exception of a timer, indicating the remaining time, which is visible only for the test group. The subjects of the study are mostly college students at both undergraduate and graduate levels. We expect our results to be fairly localized to the context of college students and their performance of standardized tests. However, this experiment can be

[1] https://webs.wofford.edu/pittmandw/psy330/exps/2011/HRexp2.htm

repeated with a larger set of assessments with varying difficulty levels and a diverse set of subjects to generalize the results to a larger population.

2. Experimental Design

The causal question of interest to us in this paper is whether the student performance is impacted by the presence of an on-screen timer. We chose to answer this question by setting up a randomized control trial with individuals as the unit of analysis. By randomizing subjects into test and control groups we expect all the factors, including unobservables, that impact the reading scores to be evenly distributed across both the groups except the treatment — presence of an on-screen timer. Therefore, using this experiment design we should be able to isolate the impact of treatment on the reading test scores. Thus, the treatment here is the on-screen timer. And because the experiment is purely randomized (all compilers), the only difference in the outcome will be because of the treatment.

We built the reading test as a survey on Qualtrics beginning with an introduction page, to set the context of the survey, followed by three board sections for data collection.

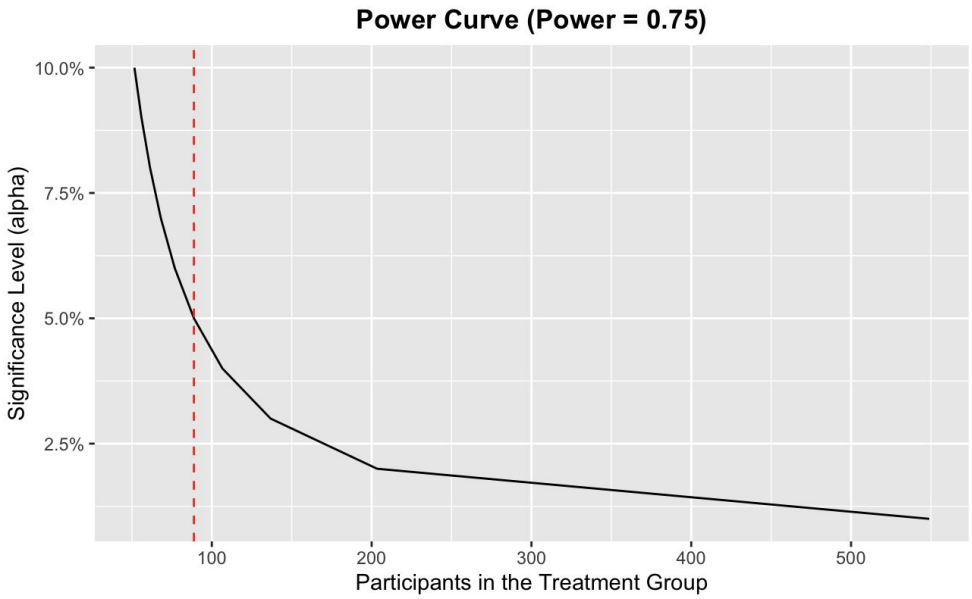
1. **Pre-test:** To ensure that the experiment has been randomized correctly, we collected demographic data (gender, age, level of education, and race of the subjects) and information about the subject's familiarity with the topic.
2. **Reading passage:** Subjects are given a GRE-level 400 words essay with a maximum reading time of 3 minutes. The treatment group has an always-on screen timer showing a countdown from 3 minutes to zero while the control group does not have this timer. After 3 minutes, the survey will auto-advance to follow-up questions (see Figure 2.1 on Appendix).
3. **Assessment:** 8 questions (multiple choices and select all that apply) to check reading comprehension, with a maximum score of 12 and minimum of 0.

We spread the survey among our circle and social channels to crowdsource the participants. The survey was active for two weeks, and once collected, we did data cleaning in Excel and R followed by analysis in R.

3. Data Collection

3.1. Power Analysis

Before we began data collection, we conducted a power test to estimate the sample size we would need for the desired lift. We set the baseline probability (p_1) to be 0.01 and the baseline plus lift probability (p_2) to be 0.1. With a modest assumption of 60 participants in the control group (n_1) and a desired power of 0.75, the following power curve follows:



Thus, we had anticipated about ~89 participants in the treatment group (n_2) given our desirable settings (total of 149). Unfortunately, the experiment only received 144 total participants, with 83 in the treatment group and 64 in the control.

Following are the variables we collected from our participants:

1. Demographic variables:
    a. Gender

    b. Age
    c. Highest level of education
    d. Race
2. Prior knowledge about:
    a. Shark conservation efforts (on the scale of 5)
    b. Shark steaks and shark-fins' use for culinary preparations
3. Post-test:
    a. Had the participant read the passage before?
4. Trackers:
    a. Survey completion rate: total percentage of survey completed
    b. Total time: total time spent on the survey
    c. Finished survey: whether or not participant successfully filled the survey
    d. Reading time: time participants spent to the read the paragraph (max 3 minutes)
    e. Question completion time: time taken to answer the reading comprehension questions

The demographic variables were voluntary, as we wanted to respect the participant's privacy and liberty to self-report their demographics. The prior knowledge asked questions about whether the participants were aware of the sharks' conservation efforts and their fins' use for culinary preparations before they were assigned a reading. These questions, along with demographics, help us assess our randomization - whether the treatment and control group were comparable "to begin with."

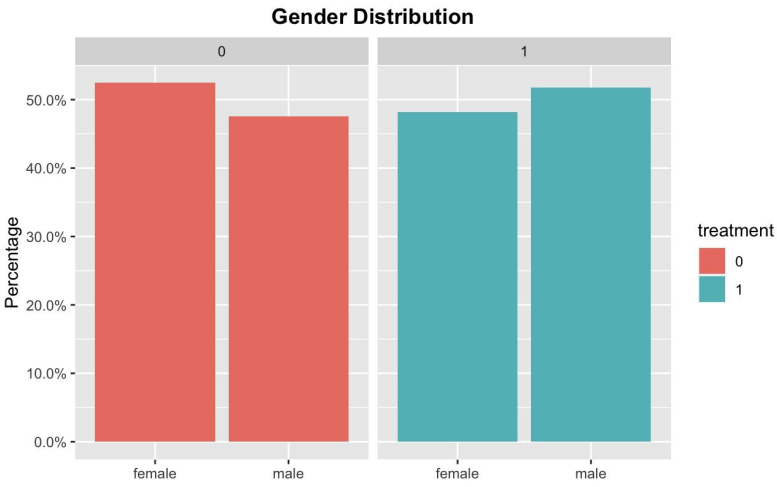
Moreover, after the reading assignment, a post-test question followed to ask whether the participants had already read the passage before. It is one of the variables that we had controlled for on our regression analysis because they will likely skew the results otherwise. Furthermore, we had trackers enabled on Qualtrics that availed us with variables mentioned under 'Trackers' above.

Finally, the participants were assigned a score (out of 12) based on the number of answers they got correct. All in all, the sample size of the study is 144 with 83 participants in the treatment group and 64 in the control. The limitations of the data collection and the results obtained are mentioned in Discussions (Section 5).

3.2 Summary Statistics

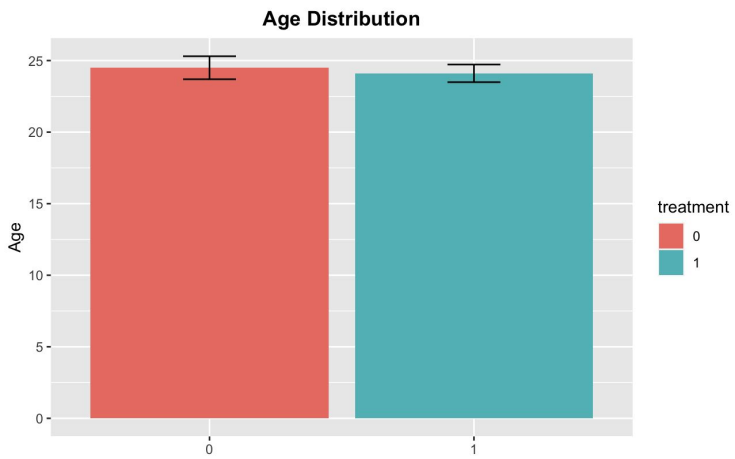
A selected number of variables below address how the treatment and the control group are similar to begin with.

3.2.1. Gender



As the bar graph of gender distribution depicts, the control group slightly had a higher female percentage (~52%) than male, and the other way around is true for the treatment group. The chi-square test (see Appendix) gives us the p-value of 0.7359, which means these two groups are not significantly different from each other.

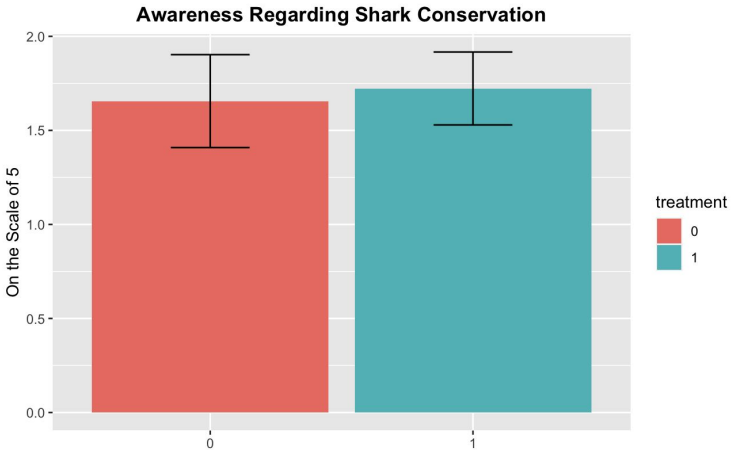
3.2.2. Age



The average age of participants in the treatment group is ~24.11 years, compared to ~24.5 years of the control group. As the overlap in error bars in the bar graph demonstrate, the age

difference among the two groups is not statistically significant.

3.2.3 Prior awareness about shark conservation

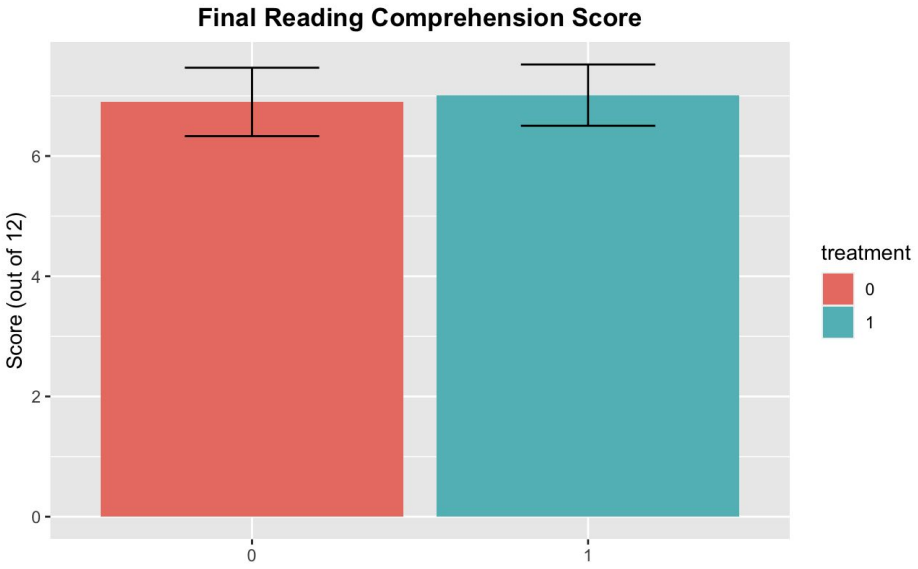


Participants in both groups were asked how well informed they were regarding the shark conservation, on the scale of 1 to 5. We can notice that the distribution is similar across two groups, and the difference is not statistically significant. This further confirms our randomization process that the control group is the counterfactual to the treatment group and hence comparable.

4. Results

Our original interest was to assess how, if at all, the timer affects the reading comprehension (RC) score. However, courtesy of Professor Pedro's suggestion, the treatment, in itself, could have spillover effects. For example, the presence of the timer might intimidate participants, prompting them to quit the survey earlier. Or, the timer could serve as a stressor that might make surveyees spend less time on the questions in the aftermath. As a result, we will measure four variables as our outcome variable of interest: (1) reading comprehension score, (2) survey completion, (3) reading time, and (4) question completion time.

4.1. Reading Comprehension Score



The figure above shows that the on-screen timer had no statistically significant difference on the final reading comprehension score. Out of 12, the treatment group had an average score of approximately 7.01, whereas the control group scored 6.9 on average. And the difference, as the overlap in the error bars depict, is not statistically significant. We also replicated this ITT result on a regression (see Regression 1 on Appendix) and found that while the participants in the treatment group scored on average 0.125 units more but that difference was not statistically significant. To further refine, we wondered if our causal effect was heterogeneous with education level as the moderator, and the treatment was not statistically significant on either of the regressions (Regression 1 on Appendix). Similarly, we also ran the heterogeneity regression with gender as the moderator, and coefficients were not significant again (refer to Regression 1.1 on Appendix).

A shortcoming in our regression here was the time participants spent to read the paragraph. Because the experiment allowed the participants to advance at their will while reading the paragraph, we had many instances where they spent as little as 2 seconds. With the designated time of 3 minutes, those participants are unlikely to have

read the passage; thus, we omitted observations in this regression where participants spent less than 60 seconds (read time < 60) on the passage.

4.2. Survey Completion Rate

The regression results of survey completion rate is mentioned in Regression 2 in Appendix. Surprisingly, with the reading time filter (read time > 60), the survey completion rate increased by 3.91 percentage points, and the increase is statistically significant at 5% level . In other words, the presence of timer increased the completion rate by 3.91 percentage points on average; however, the causal effect must be interpreted with caution as our survey is underpowered.

4.3. Passage Reading Time

Among the outcome variables of interest was how much participants spent on reading the passage. Although the designated time was 3 minutes, participants were free to advance to the RC questions at their will. This propensity could further be escalated by the timer, which could serve as a stressor. Thus, we regressed the treatment variable with the passage reading time. And with the reading time filter like in 4.1 and 4.2, the treatment of on-screen timer increased the reading time by 13.88 seconds on average, and the difference is statistically significant at 10% level. In other words, participants in the treatment group spent 13.88 more seconds on average than their counterparts in the control group (see Regression 3 in Appendix).

4.4. Question Completion Time

Similar to 4.3, the visibility of timer could also potentially affect how much time participants spend on answering the RC questions after reading the passage. And as the regression 4 in the Appendix demonstrates, the treated units spent 42.3 more seconds on average than did the control units, and the increase is statistically significant at 5% level. The timer, perhaps, makes readers more attentive. Something noteworthy in this regression is the filter for question completion time. The regressions above include reading time filter - the time participants spend on reading the passage.

Similar timing exists for question completion as well, meaning how much time participants spent answering the follow-up questions. A participant who spent, say 5 seconds, likely hastily clicked on the answers, or none at all. Thus, to preclude skewness, in addition to the reading time filter, we excluded observations where surveyees spent less than 60 seconds on answering the questions.

5. Discussions & Limitations

All in all, the timer did not have any statistically significant effect on the reading comprehension score. Yet, because of the timer, the survey completion rate increased by 3.9 percentage points on average in the treatment group and the question completion time by 42.3 seconds - both statistically significant at 5% level. Furthermore, the treated units also spent 13.88 more seconds on average on the passage, with the statistical significance at 10% level.

As with any empirical study, our experiment also suffers from limitations - the most prominent being the power. As mentioned in 3.1, our study could not attain the desirable number of participants; thus, any resultant effect, thereof, could simply be because we do not have enough observations. An iteration of the regression with increased sample size will yield much more robust results.

Furthermore, the RC score could perhaps be affected by the difficulty of the passage itself as well. We do not have any instrument to assess its difficulty level; thus, all participants could score low (or high) regardless of the presence of the timer. Similarly, our study fails to contextualize the effect of timer with respect to allotted time for the passage. As with passage difficulty, the timer could have different effects on reading comprehension with varying time lengths. Future studies will benefit from varying levels of passage difficulty and time lengths and could assess how the timer affects those different settings.

Appendix

Figure 2.1. Screenshot of Reading Passage on Qualtrics

Growing Taste for Shark Steaks and Shark-fin Soup

02 11

A growing taste for shark steaks and shark-fin soup has, for the first time in 400 million years, put the scourge of the sea at the wrong end of the food chain. Commercial landings of this toothsome fish have doubled every year since 1986, and shark populations are plunging.

Sharks do for gentler fish what lions do for the wildebeest: they check populations by feeding on the weak. Also, sharks apparently do not get cancer and may therefore harbor clues to the nature of that disease. Finally, there is the issue of motherhood. Sharks bear their young ones alive and swimming (not sealed in eggs). Shark mothers generally give birth to litters from eight to twelve pups and bear only one litter every other year. This is why sharks have one of the lowest fertility rates in the ocean. The female cod, for example, spawns annually and lays a few million eggs at a time. If three-quarters of the cod were

Regression 1

Simple OLS Regression

	Dependent variable:	
	final_score	
	(1)	(2)
treatment1	0.125	3.504
	(0.387)	(2.261)
educationassociate	7.799***	7.910***
	(2.846)	(2.830)
educationbachelors	6.427***	9.675***
	(2.069)	(3.037)
educationhigh sch	7.248***	7.285***
	(2.215)	(2.199)
educationmasters	6.289***	9.985***
	(2.120)	(3.107)
educationphd	5.865**	10.910***
	(2.477)	(3.555)
educationsome college	6.685***	7.927***
	(2.301)	(2.399)
treatment1:educationassociate		
treatment1:educationbachelors		-3.092
		(2.324)
treatment1:educationhigh sch		
treatment1:educationmasters		-3.828
		(2.344)
treatment1:educationphd		-6.544*
		(3.476)
treatment1:educationsome college		
Constant	-1.410	-4.966
	(3.261)	(3.939)
Education Interaction Term	No	Yes
Observations	108	108
R2	0.242	0.278
Adjusted R2	0.128	0.142
Residual Std. Error	1.860 (df = 93)	1.846 (df = 90)
F Statistic	2.124** (df = 14; 93)	2.040** (df = 17; 90)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression 1.1

Simple OLS Regression

	Dependent variable:	
	final_score	
	(1)	(2)
treatment1	0.125	-0.139
	(0.387)	(0.575)
gendermale	1.123***	0.834
	(0.389)	(0.607)
treatment1:gendermale		0.493
		(0.791)
Constant	-1.410	-1.236
	(3.261)	(3.284)
Gender Interaction Term	No	Yes
Observations	108	108
R2	0.242	0.245
Adjusted R2	0.128	0.122
Residual Std. Error	1.860 (df = 93)	1.866 (df = 92)
F Statistic	2.124** (df = 14; 93)	1.995** (df = 15; 92)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression 2

	Dependent variable:	
	completion_rate	
	(1)	(2)
treatment1	2.690	3.911**
	(2.564)	(1.915)
Constant	95.498***	106.202***
	(23.667)	(16.465)
Reading Time Filter	No	Yes
Observations	143	115
R2	0.084	0.095
Adjusted R2	-0.016	-0.022
Residual Std. Error	14.430 (df = 128)	9.707 (df = 101)
F Statistic	0.837 (df = 14; 128)	0.815 (df = 13; 101)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression 3

	Dependent variable:	
	read_time	
	(1)	(2)
treatment1	7.650	13.877*
	(10.466)	(7.425)
Constant	168.238*	194.207***
	(96.590)	(63.850)
Reading Time Filter	No	Yes
Observations	143	115
R2	0.080	0.129
Adjusted R2	-0.021	0.016
Residual Std. Error	58.892 (df = 128)	37.643 (df = 101)
F Statistic	0.795 (df = 14; 128)	1.147 (df = 13; 101)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression 4

	Dependent variable:	
	time_on_questions	
	(1)	(2)
treatment1	29.313	42.342**
	(20.562)	(20.958)
Constant	-92.659	23.189
	(178.842)	(178.556)
Reading Time Filter	No	Yes
Question Completion Time Filter	No	Yes
Observations	125	105
R2	0.112	0.088
Adjusted R2	-0.001	-0.031
Residual Std. Error	106.810 (df = 110)	101.160 (df = 92)
F Statistic	0.988 (df = 14; 110)	0.742 (df = 12; 92)
Note:		*p<0.1; **p<0.05; ***p<0.01

Code

````
---
title: "A/B Testing Project"
author: "Team 2"
date: "December 18, 2020"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE)
```

```{r Import Packages}
library(tidyverse)
library(dplyr)
library(ggplot2)
library(cobalt)
library(scales)
library(data.table)
library(stargazer)
library(plm)
library(usdm)
library(pwr)
library(ggthemes)
```

```{r Import Dataframe}
df <- read_csv("project_data_cleaned.csv")
```


```{r Data Cleaning 1}
df$finished <- tolower(df$finished)
df$education <- as.character(df$education)
df$education[df$education == "Master's degree"] <- "masters"
df$education[df$education == "Less than high school degree"] <- '< high sch'
df$education[df$education == "Bachelor's degree in college (4-year)"] <- 'bachelors'
df$education[df$education == "Doctoral degree"] <- 'phd'
````
````

```
df$education[df$education == "High school graduate (high school diploma or equivalent
including GED)"] <- "high sch"
df$education[df$education == "Some college but no degree"] <- 'some college'
df$education[df$education == "Associate degree in college (2-year)"] <- 'associate'

df$education <- as.factor(df$education)
```

```{r Data Cleaning 2}
df_new <- dplyr::select(df,completion_rate, total_time, finished, gender, age,
education, race,
                Q2, Q3, treatment, treatment_page_submit, control_page_submit,
                rc_page_submit, Q17, score_final)

names(df_new) <- c("completion_rate","total_duration","finished", "gender","age",
"education","race","awareness_shark", "awareness_shark_food",
"treatment","read_time_treat","read_time_control","time_on_questions","post_test","fin
al_score")

df_new$treatment <- as.factor(as.character(df_new$treatment)) # treatment variables as
a factor

#removing entries where treatment is N/A
df_final <- df_new[!is.na(df_new$treatment), ]
df_final <- df_final[df_final$treatment != "#N/A",]

df_final$finished <- with(df_final, ifelse(completion_rate==100, "true", "false"))

#combining the reading time of treatment and control group in a single column
df_final$read_time_treat <- with(df_final, ifelse(is.na(read_time_treat),
read_time_control, read_time_treat))

names(df_final)[names(df_final) == "read_time_treat"] <- "read_time"

#drop the reading time of both the treatment and the control group as they are already
combined now
dropcols <- c("read_time_control")
df_final<-df_final[ , !(names(df_final) %in% dropcols)]
```

```
#age variable as an integer
df_final$age <- as.integer(as.character(df_final$age))
df_final$treatment <- factor(df_final$treatment)
```

The total number of observations: `r nrow(df_final)`, with `r nrow(df_final[df_final$treatment ==1,])` participants in the treatment group and `r nrow(df_final[df_final$treatment ==0,])` in the control group.

Exploratory Data Analysis

```{r Survey Completion}
df_final %>%
  ggplot(aes(x = finished, fill = treatment))+
  geom_bar(aes(y = ..prop.., position = "dodge", group = 1)) +labs(y = "Percentage", x = NULL)+ ggtitle("Survey Completion of Participants") + theme(plot.title = element_text(hjust = 0.5, face = "bold")) + facet_wrap(.~treatment) + scale_y_continuous(labels = percent)
```

```{r Gender Distribution}
df_final %>%
  ggplot(aes(x = gender, fill = treatment)) +
  geom_bar(aes(y = ..prop.., position = "dodge", group = 1)) +
  labs(y = "Percentage", x = NULL) +
  ggtitle("Gender Distribution") + theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  facet_wrap(.~treatment) + scale_y_continuous(labels = percent)
```

```{r Education Level, fig.height =10}
df_final %>%
  ggplot(aes(x = education, fill = treatment))+
  geom_bar(aes(y = ..prop.., position = "dodge", group = 1)) +
  labs(y = "Percentage", x = NULL) +
  ggtitle("Education Distribution") + theme(plot.title = element_text(hjust = 0.5)) +
  facet_wrap(.~treatment) +
  theme(axis.text.x = element_text(angle = 90, hjust =1, vjust =1, size =9)) +
```

```
  scale_y_continuous(labels = percent)
```

```{r Chi-Squared Test}
chisq.test(df_final$gender, df_final$treatment)
chisq.test(df_final$race, df_final$treatment)
chisq.test(df_final$post_test, df_final$treatment)
chisq.test(df_final$education, df_final$treatment)
```

```{r Pre-test 1: Awareness Regarding Shark Conservation}
summary <- df_final[!is.na(df_final$awareness_shark),] %>%
  group_by(treatment) %>%
  summarise(count = n(), avg.diff = mean(awareness_shark),
            lower = t.test(awareness_shark, mu = 0)$conf.int[1],
            upper = t.test(awareness_shark, mu = 0)$conf.int[2])

summary$treatment <- as.factor(as.character(summary$treatment))

summary %>%
  ggplot(aes( x= treatment, y= avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.3,
                position = position_dodge(0.9)) +
                labs(y = "On the Scale of 5",
                     x = NULL) +
  ggtitle("Awareness Regarding Shark Conservation") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```{r}
summary <- df_final[!is.na(df_final$age),] %>%
  group_by(treatment) %>%
  summarise( count = n(), avg.diff= mean(age), lower = t.test(age, mu =0)$conf.int[1],
upper = t.test(age, mu =0)$conf.int[2])

summary %>%
  ggplot(aes( x= treatment, y= avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
```

```
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.2,
                position = position_dodge(0.9)) + labs(y = "Age", x = NULL) +
  ggtitle("Age Distribution") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```{r Pre-test 3: Awarness Regarding Shark Food}
summary <- df_final %>%
  group_by(awareness_shark_food, treatment) %>%
  summarise(count = n())

df_final %>%
  ggplot(aes(x = awareness_shark_food, fill = treatment)) +
  geom_bar(aes(y = ..prop.., position = "dodge", group = 1)) +
  labs(y = "Percentage", x = NULL) +
  ggtitle("Prior Awareness about Shark Food") +
  facet_wrap(.~treatment) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  scale_y_continuous(labels = percent)
```

```{r}
summary <- df_final %>%
  group_by(treatment) %>%
  summarise(count = n(), avg.diff= mean(total_duration/60),
            lower = t.test(total_duration/60, mu=0)$conf.int[1],
            upper = t.test(total_duration/60, mu =0)$conf.int[2])

summary %>%
  ggplot(aes( x= treatment, y = avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.4,
                position = position_dodge(0.9)) +
  labs(y = "Total Time (in minutes)", x = NULL) +
  ggtitle("Total Survey Duration Distribution") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

```

```{r}
summary <- df_final[!is.na(df_final$read_time),] %>%
  group_by(treatment) %>%
  summarise(count = n(), avg.diff= mean(read_time),
            lower = t.test(read_time, mu =0)$conf.int[1],
            upper = t.test(read_time, mu=0)$conf.int[2])


summary %>%
  ggplot(aes( x= treatment, y= avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.4,
                position = position_dodge(0.9)) +
  labs(y = "Total Time (in seconds)", x = NULL)+
  ggtitle("Total Reading Time Distribution") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```{r}
summary <- df_final[!is.na(df_final$time_on_questions),] %>%
  group_by(treatment) %>%
  summarise(count = n(), avg.diff= mean(time_on_questions),
            lower = t.test(time_on_questions, mu =0)$conf.int[1],
            upper = t.test(time_on_questions, mu =0)$conf.int[2])

summary %>%
  ggplot(aes(x = treatment, y= avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.4,
                position = position_dodge(0.9)) +
  labs(y = "Total Time (in seconds)", x = NULL) +
  ggtitle("Time Spent on Questions") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```{r}
summary <- df_final[!is.na(df_final$final_score),] %>%
```

```
  group_by(treatment) %>%
  summarise(count = n(), avg.diff= mean(final_score),
            lower = t.test(final_score, mu=0)$conf.int[1],
            upper = t.test(final_score, mu =0)$conf.int[2])

summary %>%
  ggplot(aes( x= treatment, y= avg.diff, fill = treatment)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper),
                width =0.4,
                position = position_dodge(0.9)) +
  labs(y = "Score (out of 12)", x = NULL) +
  ggtitle("Final Reading Comprehension Score") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```{r}
df_test <- df_final[,c("age", "gender","finished","education","race",
"awareness_shark_food", "post_test", "treatment" )]
df_test <- na.omit(df_test)

df_test$treatment <- as.integer((as.character(df_test$treatment)))
df_test$p.score <- glm(f.build("treatment", c("gender", "race", "education")),
                       data = df_test, family = "binomial")$fitted.values
df_test$att.weights <- with(df_test, treatment + (1-treatment)*p.score/(1-p.score))

bal.tab(treatment ~ gender, data = df_test, weights = "att.weights", distance =
"p.score")
```

Results Obtained

```{r Power Analysis}
#power analysis

#baseline probability and lift required
P1 = 0.01 # baseline probability
P2 = 0.1 # baseline+lift probability
```

```
#plot power curve
slevels<-seq(from=0.01, to=0.10, by=0.01)
ptab <- cbind(NULL, NULL)

for (i in 1:length(slevels)) {
pwrt <- pwr.t2n.test(ES.h(P1,P2),
                     n1 = 60,
                     n2 = NULL,
                     sig.level = slevels[i],
                     power = 0.75,
                     alternative="two.sided")
ptab <- rbind(ptab, cbind(pwrt$n2, pwrt$sig.level))
}

temp <- ptab %>% as.data.frame()
ggplot(temp, aes(x = V1, y = V2)) +
  geom_line() +
  geom_vline(xintercept = temp$V1[temp$V2 == 0.05], linetype = "dashed", color =
"red", size = 0.5) +
  labs(y = 'Significance Level (alpha)', x = 'Participants in the Treatment Group') +
  ggtitle("Power Curve (Power = 0.75)") + theme(plot.title = element_text(hjust = 0.5,
face = "bold")) +
  scale_y_continuous(labels = percent)
```

```{r}
df_final$gender <- as.factor(df_final$gender)
df_final$education <- as.factor(df_final$education)
df_final$race <- as.factor(df_final$race)
df_final$post_test <- as.factor(df_final$post_test)
```

```{r Regression 1: Reading Comprehension Score}
# Regression 1: Final score as DV(first regression without interaction terms, and the
second with regression)

lm_score <- lm(final_score ~  treatment + gender + age + education + race + post_test,
data = subset(df_final, read_time > 60))
```

```
#heterogeneous causality with education level

lm_score_with_interaction <- lm(final_score ~  treatment + gender + treatment *
education +
                                  age + education + race +
                                  post_test, data = subset(df_final, read_time > 60))

stargazer(lm_score, lm_score_with_interaction,
          se = NULL,
          title = "Simple OLS Regression",
          type="text",
          omit = c("gender", "age", "race", "post_test", "read_time"),
          add.lines = list (c("Education Interaction Term","No","Yes")))
```

```{r}
#heterogeneous causality with gender

lm_score_gender <- lm(final_score ~  treatment + gender + treatment * gender +
                                  age + education + race +
                                  post_test, data = subset(df_final, read_time > 60))


#heterogeneous causality with gender

stargazer(lm_score, lm_score_gender,
          se = NULL,
          title = "Simple OLS Regression",
          type="text",
          omit = c("age", "race", "education", "post_test", "read_time"),
          add.lines = list (c("Gender Interaction Term","No","Yes")))
```

```{r Regression 2: Survey Completion Rate}

lm_completion <- lm(completion_rate ~  treatment  + gender  + age + education + race,
                    data = df_final)

lm_completion_with_filter <- lm(completion_rate ~  treatment  + gender  + age +
education + race, data = subset(df_final, read_time > 60))
```

```
stargazer(lm_completion, lm_completion_with_filter,
          add.lines = list (c("Reading Time Filter","No","Yes")),
          omit = c("gender", "age", "race", "education", "post_test", "read_time"),
          type="text")
```

```{r Regression 3: Read Time}

lm_read_time <- lm(read_time ~  treatment  + gender  + age + education + race,
                   data = df_final)

lm_read_time_with_filter <- lm(read_time ~  treatment  + gender  + age + education +
race, data = subset(df_final, read_time > 60))

stargazer(lm_read_time, lm_read_time_with_filter,
          add.lines = list (c("Reading Time Filter","No","Yes")),
          omit = c("gender", "age", "race", "education", "post_test", "read_time"),
          type="text")
```

```{r Regression 4: Question Completion Time}
lm_q_completion <- lm(time_on_questions ~  treatment  + gender  + age + education +
race,
                   data = df_final)

lm_q_completion_filter <- lm(time_on_questions ~  treatment  + gender  + age +
education + race,
                   data = subset(df_final, read_time > 60 & time_on_questions > 60))

stargazer(lm_q_completion, lm_q_completion_filter,
          add.lines = list (c("Reading Time Filter","No","Yes"),
                            c("Question Completion Time Filter","No","Yes")),
          omit = c("gender", "age", "race", "education", "post_test", "read_time"),
          type="text")
```