Speaking Science

W241 Final Project

Aris Fotkatžikis, Haerang Lee, Mumin Khan

Contents

A total of 264 people clicked on our survey on Mechanical Turk. 64 people did not complete the survey. Out of the 200 who completed the survey, we dropped 6 people, either due to missing values or because they stated that they were not LA residents. There were 40 people that their IP address placed outside CA and LA, but since they might have been people on travel we decided to keep them in our survey. We ended up having 97 people in treatment and 97 people in control. A graphical breakdown of ourparticipants is shown below.

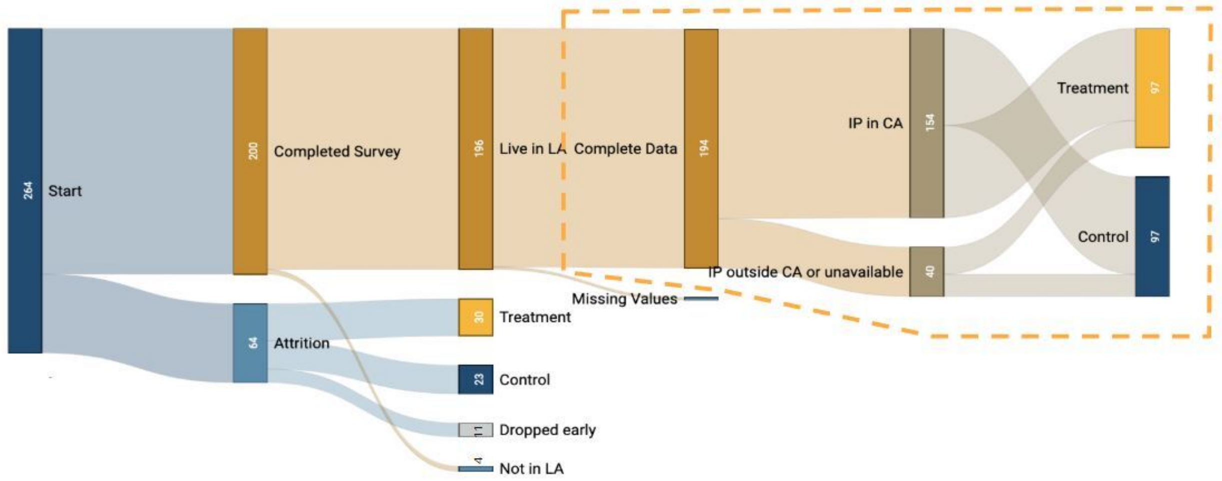


Figure 1: Sankey Diagram of survey data

Let's import the data for the 196 participants that comprise our treatment and control groups.

```
##  response_id        duration_in_seconds time_read_article  credibility
##  Length:194         Min.   :  76.0      Min.   : 16.04    Min.   :2.000
##  Class :character   1st Qu.: 245.5      1st Qu.: 48.76    1st Qu.:5.000
##  Mode  :character   Median : 391.5      Median :155.92    Median :6.000
##                     Mean   : 430.7      Mean   :190.44    Mean   :5.758
##                     3rd Qu.: 597.8      3rd Qu.:286.76    3rd Qu.:6.000
##                     Max.   :1586.0      Max.   :743.89    Max.   :7.000
##    importance      q1_correct       q2_correct      q3_correct
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:6.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :6.000   Median :1.0000   Median :0.000   Median :0.0000
##  Mean   :6.062   Mean   :0.5258   Mean   :0.299   Mean   :0.3918
##  3rd Qu.:7.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :7.000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##    q4_correct       q5_correct       q6_correct      questions_correct
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.0000   Median :1.0000   Median :1.0000   Median :3.000
##  Mean   :0.1753   Mean   :0.5619   Mean   :0.6082   Mean   :2.562
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :5.000
##  time_answering_questions    donation        treatment
##  Min.   :   8.679         Min.   :  0.00   Min.   :0.0
##  1st Qu.:  76.507         1st Qu.:  0.00   1st Qu.:0.0
##  Median : 118.156         Median : 17.50   Median :0.5
##  Mean   : 149.991         Mean   : 29.13   Mean   :0.5
##  3rd Qu.: 177.875         3rd Qu.: 50.00   3rd Qu.:1.0
##  Max.   :1091.081         Max.   :100.00   Max.   :1.0
```

We see that there are no missing values and nothing appears out of order. We examined many aspects of our data. In this EDA portion of our report, we will only highlight key aspects of our data, to adhere to the 20 page limitation. To ensure that we could use the quiz questions to assess how engaged our survey takers were, we selected relatively dificult questions. As we can see from the summary table above, nobody got all 6 questions correct. The distribution of the number of correct answers is shown below.This distribution is quite similar to the one we expected based on our pilot study (**WE NEED TO ADD GRAPHS OR REFER TO HAERANG'S SECTION HERE**)

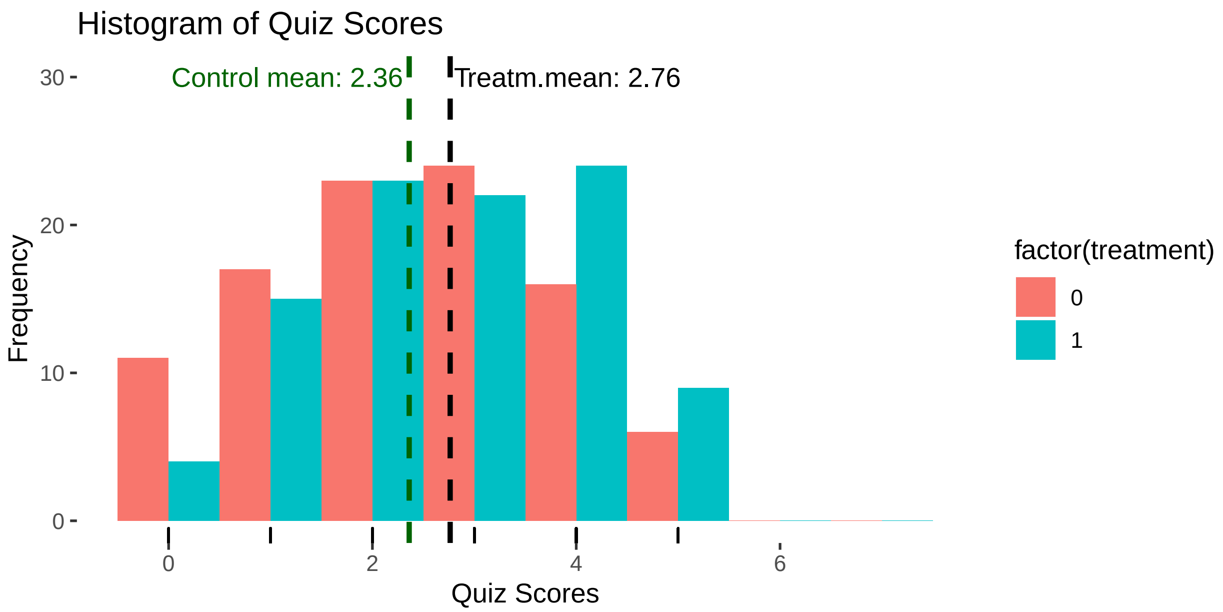
```
# Average number of correct answers per group
d[d$treatment == 0][ , mean(questions_correct)]
```

```
## [1] 2.360825
```

```
d[d$treatment == 1][ , mean(questions_correct)]
```

```
## [1] 2.762887
```

```
ggplot(d, aes(x = questions_correct, fill = factor(treatment))) +
  # geom_density(alpha = 0.4) +
  geom_histogram(position="dodge", breaks=c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5))+
  geom_vline(aes(xintercept = mean(d[d$treatment == 0, questions_correct])), color = "darkgreen", linet
  geom_vline(aes(xintercept = mean(d[d$treatment == 1, questions_correct])), color = "black", linetype =
  annotate("text", x = 2.3, y = 30, label = "Control mean: 2.36", color = "darkgreen", hjust="right") +
  annotate("text", x = 2.8, y = 30, label = "Treatm.mean: 2.76", color = "black", hjust="left") +
  ggtitle("Histogram of Quiz Scores") + geom_rug() +
  ylab("Frequency") +
  xlab("Quiz Scores")+
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())
```



We can see that Questions 4, 2, and 3 are the most difficult ones (in order of difficulty) and only 3 out of 6 questions had over 50% correct answers.

```
# Most dificult question
cat("Number of correct answers for question Q_1, is: ", sum(d$q1_correct), "\n")
```

```
## Number of correct answers for question Q_1, is:  102
```

```
cat("Number of correct answers for question Q_2, is: ", sum(d$q2_correct),  "\n")
```

```
## Number of correct answers for question Q_2, is:  58
```

```
cat("Number of correct answers for question Q_3, is: ", sum(d$q3_correct), "\n")
```

```
## Number of correct answers for question Q_3, is:  76
```

```
cat("Number of correct answers for question Q_4, is: ", sum(d$q4_correct), "\n")
```

```
## Number of correct answers for question Q_4, is:  34
```

```
cat("Number of correct answers for question Q_5, is: ", sum(d$q5_correct), "\n")
```

```
## Number of correct answers for question Q_5, is:  109
```

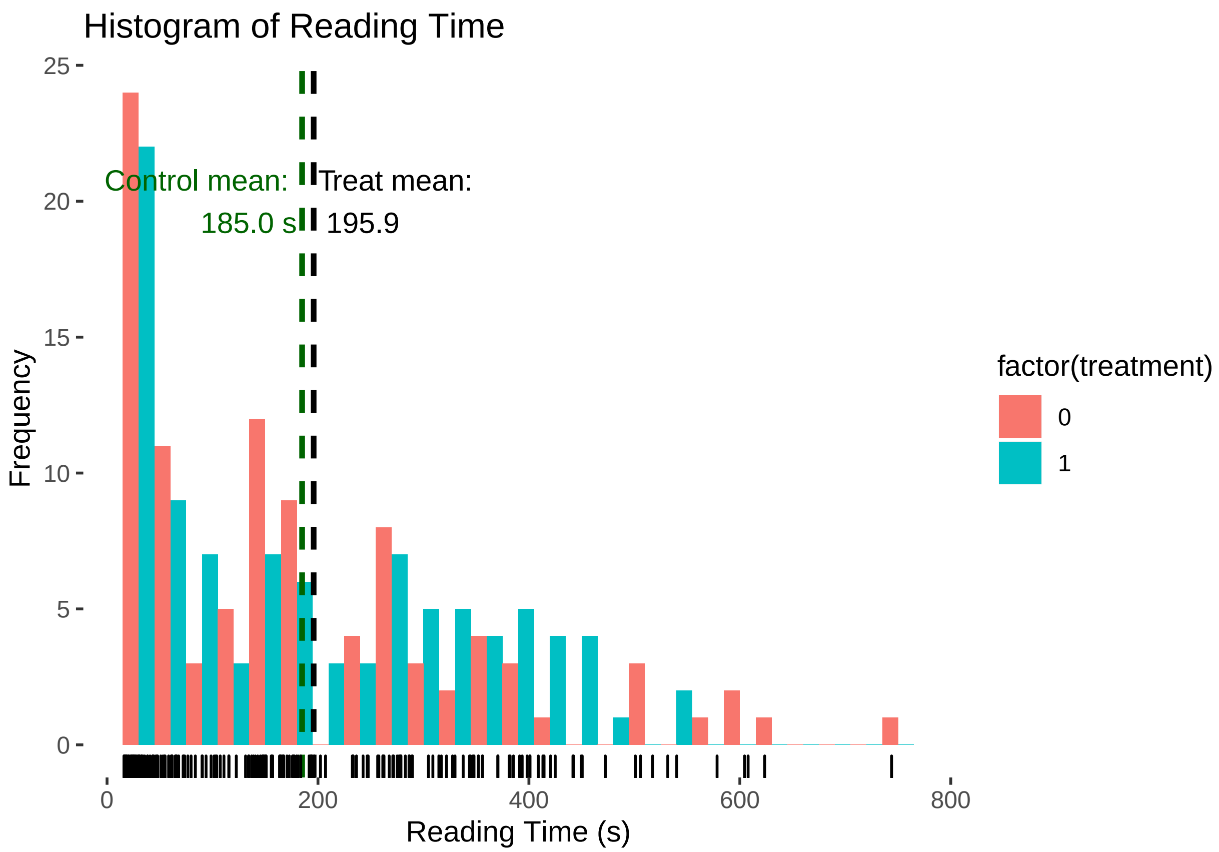
```
cat("Number of correct answers for question Q_6, is: ", sum(d$q6_correct))
```

```
## Number of correct answers for question Q_6, is:  118
```

We also check the distribution of reading times for treatment and control. We can see in the graphs below that they don't follow a normal distribution. In addition, the actual distribution of donation amounts, seems to have a much smaller difference between treatment and control than either one of our assumed distributions, based on our pilot study (**WE NEED TO ADD GRAPHS OR REFER TO HAERANG'S SECTION HERE**). Our power analysis indicated that we needed more than 600 observations, but we only had 194.

For an even smaller treatment effect, maybe we needed an even greater sample size. So we're not confident we will detect any treatment effects here.

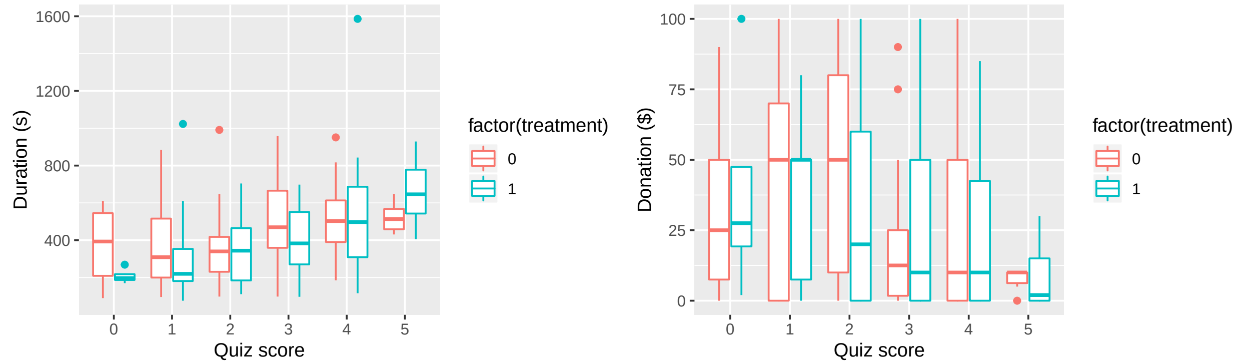
```
ggplot(d, aes(x = time_read_article, fill = factor(treatment))) +
  # geom_density(alpha = 0.4) +
  geom_histogram(position="dodge", binwidth=30)+
  geom_vline(aes(xintercept = mean(d[d$treatment == 0, time_read_article])), color = "darkgreen", linety
  geom_vline(aes(xintercept = mean(d[d$treatment == 1, time_read_article])), color = "black", linetype =
  annotate("text", x = 180, y = 20, label = "Control mean: \n 185.0 s", color = "darkgreen", hjust="righ
  annotate("text", x = 200, y = 20, label = "Treat mean:\n 195.9", color = "black", hjust="left") +
  ggtitle("Histogram of Reading Time") + geom_rug() +
  ylab("Frequency") +
  xlab("Reading Time (s)")+
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())
```



Let's check below the distribution of the survey duration and the distribution of the doonation with the quiz score (i.e. some of correct answers per survey taker). On the left we are showing the distribution of time spent reading the article with the number of correct answers. It is not surprising that in general the longer someone spent reading the article, the better they did on the quiz. This holds for treatment and control. The graph on the right, shows the distribution of the amount donated with the number of correct answers. It seems that the more correct answers a person had, the less amount of money they donated. This holds both for treatment and control. That might be an indication that the more effort people put in reading and answering the questions, the less they were inclined to donate their hard earned money.

```
a <- ggplot(d, aes(x = as.factor(questions_correct), y = duration_in_seconds, color = factor(treatment)
  xlab("Quiz score")
b <- ggplot(d, aes(x = factor(questions_correct), y = donation, color = factor(treatment) )) + geom_box
  xlab("Quiz score")

grid.arrange(a, b, ncol = 2)
```

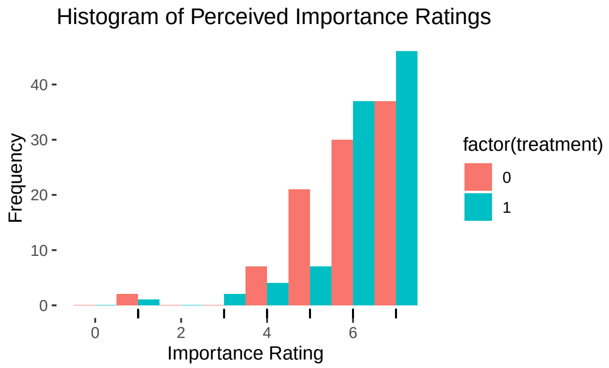


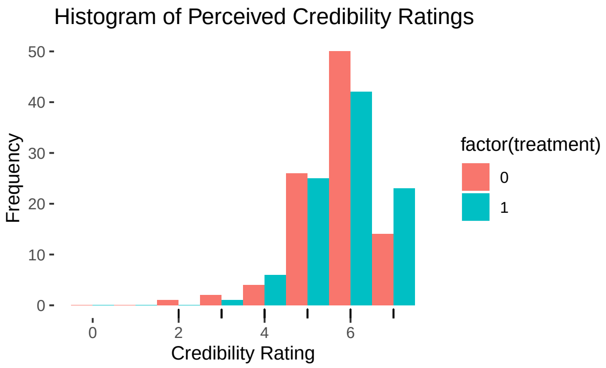
We asked survey takers to rate the Importance of the topic discussed in the article and also rate their perceived Credibility of the article on a scale from 1 to 7. Plotting the distributions of the Importance and Credibility variables, we see on the right that the credibility ratings were rather similar for both treatment and control. On the left we see that the treatment group tends to assign higher Importance scores compared to control, and this difference is statistically significant at the 0.05 level, in accordance with our hypothesis.

```
plot1 <- ggplot(d, aes(x = importance, fill = factor(treatment))) +
  # geom_density(alpha = 0.4) +
  geom_histogram(position="dodge",breaks=c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5))+
  ggtitle("Histogram of Perceived Importance Ratings") + geom_rug() +
  ylab("Frequency") +
  xlab("Importance Rating")+
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())

plot2 <- ggplot(d, aes(x = credibility, fill = factor(treatment))) +
  # geom_density(alpha = 0.4) +
  geom_histogram(position="dodge",breaks=c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5))+
  ggtitle("Histogram of Perceived Credibility Ratings") + geom_rug() +
  ylab("Frequency") +
  xlab("Credibility Rating")+
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())

grid.arrange(plot1, plot2, ncol = 2)
```





```
# Test how different the distributions for treatment and control are for the Importance and Credibility

wilcox.test(d$importance[d$treatment == 0], d$importance[d$treatment == 1])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d$importance[d$treatment == 0] and d$importance[d$treatment == 1]
## W = 3969.5, p-value = 0.04491
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(d$credibility[d$treatment == 0], d$credibility[d$treatment == 1])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d$credibility[d$treatment == 0] and d$credibility[d$treatment == 1]
## W = 4371, p-value = 0.3607
## alternative hypothesis: true location shift is not equal to 0
```

Results

Compliance & Attrition

For our final results, we've opted to not include observations for those who took fewer than 100 seconds to complete the questions portions of the survey. This is based on a question section length of 340 words and a read and answer composite time of 200 words per minute. The result is a dataset of 111 observations, 61 in control and 50 in treatment. We decided not to filter any of the Article Read Time values because our survey had forced people to stay on the page for 15 seconds. Furthermore, we believe that applying a words read per minute threshold might not be an accurate model of how people interact with journalistic writings, especially those that are scientifically oriented. We had a large number of attritors after the Mechanical Turk task was filled. We believe that this attrited was comprised of another random sampling of the population we sampled while the task was active, thus we have excluded those incomplete responses from this analysis with the belief that the exclusion won't bias our results in any direction.

Regression Results

A simple regression of our three outcome variables yields the following table.

Comparing Treatment Effects

```
                                                 Dependent variable:
```

```
-----------------------------------------------------------
Questions Correct Article Read Time (seconds) Donation in USD
      (1)                    (2)                   (3)
```

Treatment 0.591** 34.217 -4.050 p = 0.024 p = 0.276 p = 0.464

Observations 111 111 111 R2 0.046 0.011 0.005 Adjusted R2 0.037
0.002 -0.004 Residual Std. Error (df = 109) 1.352 163.609 28.886 F Statistic (df = 1; 109) 5.261** 1.202 0.540
==

Note: *p<0.1;* ***p<0.05;*** p<0.01

As shown, we observed a treatment effect of 0.5915 with a p-value of 0.0237 for the number of questions the survey taker answered correctly when treated. This hints that respondents who received the local (Los Angeles) article paid more attention to its contents and were able to recall information better on the quiz. Unfortunately, the same cannot be said about our Article Read Time outcome variable (ATE = 34.2172, p=0.2754) or our Donation ammount outcome variable (ATE = -4.0502, p=0.4639).

When considering the effect of treatment on the number of questions a respondent answered correctly, we wanted to make sure there were no unobserved confounds contributing to the effect. After running serveral analysis, the only significant covariate we found was Article read time. Taking Article Read Time into account yields the regression below.

Comparing Treatment Effects
==

Dependent variable: ———————— Questions Correct

Treatment 0.501**
p = 0.045

Article Read Time (seconds) 0.003***
p = 0.001

Observations 111 R2 0.144 Adjusted R2 0.128 Residual Std. Error 1.286 (df = 108) F Statistic 9.099*** (df = 2; 108)
==

Note: *p<0.1;* ***p<0.05;*** p<0.01

The table does show a small effect of Article Read Time on the Questions Correct outcome variable, but the treatment effect is still there at 0.5008 and is still statistically significant at the 95% conficence level (p = 0.0176).

We were not able to measure a statistically signigcant effect from either Article Time Read or Donation Ammount outcome variables. To see if we were asking the right questions, we created a binned catagory of Article Read Times for each minute and a dummy variable to represent whether or not the respondent donated. As you can see from the table below, the results are inconclusive. Log transformations of both did not help.

Comparing Treatment Effects
==

Dependent variable: ————————————————- Article Read Time (1 Minute bins) Donation > 0 (1) (2)

Treatment 0.719 -0.138
(0.523) (0.089)

Observations 111 111
R2 0.017 0.021

Adjusted R2 0.008 0.012
Residual Std. Error (df = 109) 2.739 0.467
F Statistic (df = 1; 109) 1.895 2.386

==

Note: *p<0.1;* ***p<0.05;*** p<0.01

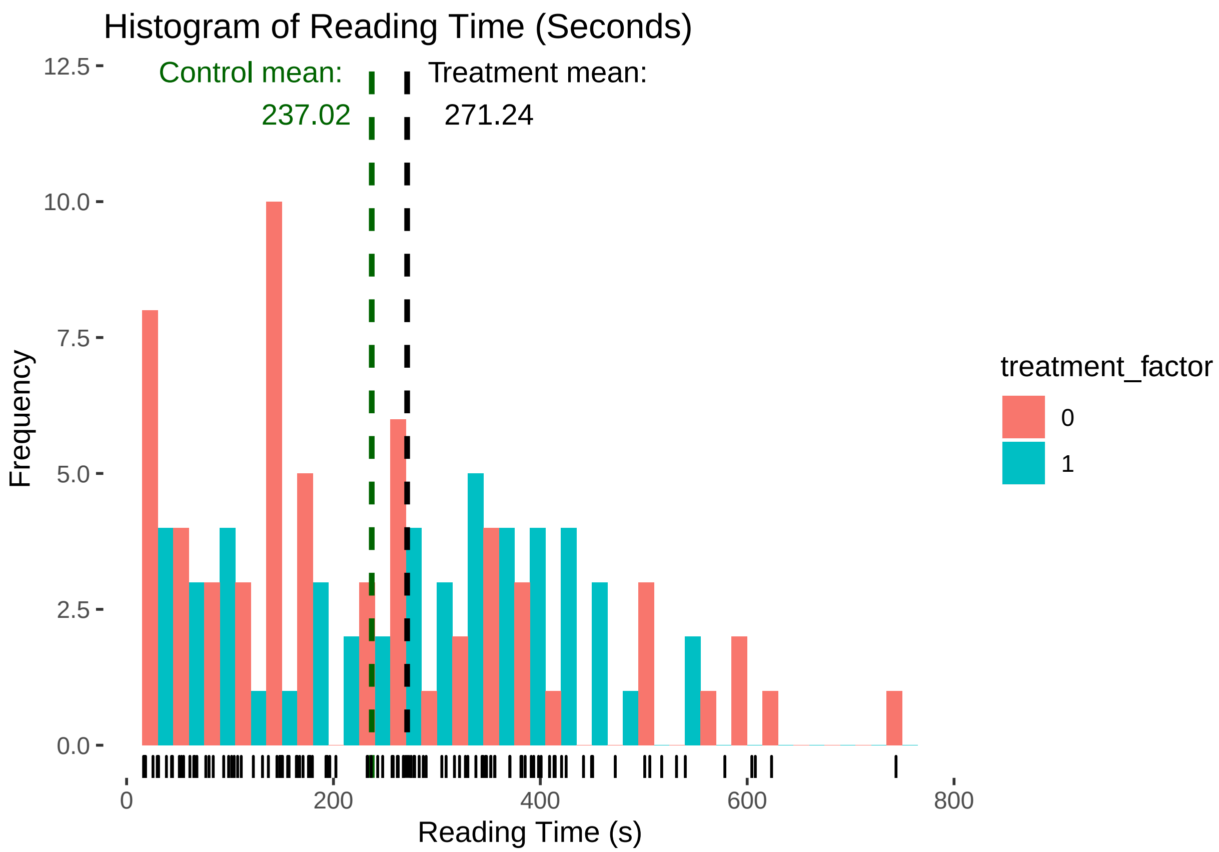
Compliance

A surprising finding of our survey was how little time was spent reading the article. We had initially anticipated a normal or normal-esque shaped distribution of reading times around the 2 minute mark. Instead, we observed a highly right-skewed distribution of article read times.

```r
d$treatment_factor <-as.factor(d$treatment)

mean_control_time <- round(mean(d[d$treatment_factor == 0, time_read_article]), 2)
mean_treatment_time <- round(mean(d[d$treatment_factor == 1, time_read_article]), 2)


ggplot(d, aes(x = time_read_article, fill = treatment_factor)) +
  # geom_density(alpha = 0.4) +
  geom_histogram(position="dodge", binwidth=30)+
  geom_vline(aes(xintercept = mean_control_time), color = "darkgreen", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = mean_treatment_time), color = "black", linetype = "dashed", size = 1) +
  annotate("text", x = mean_control_time - 20, y = 12, label = paste("Control mean: \n ", mean_control_
  annotate("text", x = mean_treatment_time + 20, y = 12, label = paste("Treatment mean: \n ", mean_trea
  ggtitle("Histogram of Reading Time (Seconds)") + geom_rug() +
  ylab("Frequency") +
  xlab("Reading Time (s)")+
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())
```



Generalizability

Science Communication is Broad

Comprehension is Difficult to Quantify

Competing Incentives with Mechanical Turk

Conclusion