Speaking Science

W241 Final Project

Aris Fotkatžikis, Haerang Lee, Mumin Khan

Contents

Executive Summary

In this experiment (N = **196** *[HL: check final N after exclusion]*), we examine whether the proximity to the location discussed in a science news article affects the reader's engagement with the article topic. We found that _____

Introduction

Well-written science news spurs interest in the topic and inspires actions. One might expect a well-informed article on diabetes to compel readers to eat healthier. The news about the wonders of Mother Nature may inspire children to pay closer attention to the trees when they hike, to visit museums more, or perhaps even pursue a career in science. We started with the question of how effective science communication could increase the public's engagement with science.

While an examination of long-term engagements would have been ideal, we were strapped for resources. Then, we remembered a recent news article about how the rising sea levels may change the landscape of the Bay Area, where we live. It had compelled us to research the topic and become knowledgeable about how to change our behaviors to alleviate climate change. For this study, our research question focuses instead on the very first step of the long-term engagement: **Does the locality of a scientific article impact the reader's engagement with the issue discussed in the article?**

We recruited survey respondents from Los Angeles. We showed them an article that discussed air pollution caused by commercial ships visiting the regional ports. For those in the treatment group, the article referred to Los Angeles, and for those in the control group, the article referred to New York.

After they had read the articles, we asked them questions to collect the following outcomes.

1. Reading comprehension quiz score (0 - 6 points)
2. Donation amount toward alleviating air pollution, if they won a $100 raffle from this survey
3. Article reading time
4. Rating of importance of the air pollution issue
5. Rating of credibility of the article

We hypothesize that a locally-related topic would make people read the article more carefully and understand more of it. The implications of the results would mean that the science communities that hope to engage more people should target local communities.

Experimental Design

Narrowing Down the Question

The most challenging part of our journey was to crystallize the research question. We could have operationalized the concept of engagement in myriad ways: future college majors in a scientific discipline; museum visits; the likelihood of creating a vinegar volcano with kids at home; or eco-friendly product consumption.

In the first iteration of our research design, we considered testing whether the jargon and the author of a tweet would change the click-through rate (CTR). However, we did not believe that clicking on a link necessarily indicated an intrinsic fascination with science or signaled a long-term effect.

We reached out to a contact at the Lawrence Berkeley National Laboratory called Tim Hurt, who develops science education curriculums. He shared some surveys that his team used to measure students' interest or fascination with science. We considered using a modified version of this survey, which would have given us a self-reported level of engagement. However, we did not believe that a feeling of engagement with science would change within the short timeframe of our research. Finally, only after spending a long time exploring two utterly different research questions, we arrived at the idea of measuring the impact of local news on engagement.

Treatment

The treatment is the locality of the article. We originally wanted to compare the effect of an article set in the reader's city, compared to a broader region like the country or even the world. The comparison between a city and a broader region may have been possible if we could have a bank of news articles, from which we could randomly select and give to each survey respondent. Such a design would require a large number of articles to control for other variables that may affect the outcome, including the source, topic, and diction.

Due to the limited resources we had, we wanted to stick to a single article in which the location keywords would be the only difference between treatment and control. Unfortunately, we could not exchange the name of a city for a state, country, or the world without completely changing the contents of the sentence and the article. So we changed the control variable to represent a remote city.

(Seattle consideration here)

Choosing the Article

We gave careful considerations to the topic, which could influence the outcome. For instance, a divisive topic such as climate change could impact the engagement of a subset of the readers with a particular political affiliation and violate the excludability assumption. On the other extreme, a highly esoteric topic could fail to engage a majority of our readers. After some discussion, we determined that pollution and its impact on people's lives could be neutral and relatable enough for the study.

In our pilot study, we discovered more variables that affected the outcome. All of the respondents scored 100% on the reading comprehension quiz, because the article was too short and easy. The lack of variance implied that any existing treatment effect would be hard to detect. So we picked a new article that was longer and made the quiz harder.

Our article of choice was about large ships contributing to air pollution. We recruited participants from LA and used Qualtrics features to randomly assign treatment by the individual. The treatment group was given the article with references to LA. In contrast, the control group was given the same article with references to New York, because New York was a coastal city with a similar population as LA and had large ports that the article could be about.

Power Analysis & Outcome

Reading Comprehension

We hypothesized that if a participant finds an article more interesting because it is related to their immediate residential environment, then they would read it more closely and retain more of the information. To measure the information retention, we designed a 6-question quiz and used quiz scores as an outcome variable representing engagement, fully recognizing the following limitations to this measurement. First, we assume that the quiz performance is an accurate representation of the comprehension. Second, we treat the score as a continuous variable, although one's comprehension may not be on a linear scale.

After conducting a pilot study, we studied the results and determined that a distribution like the following may be realistic to expect from the full study.

(graph of assumed distribution)

Assuming that the above represents the true distribution, the power analysis indicated that we would need at least 100 respondents to achieve a power of 90%.

(insert power graph)

Donation as an Action

We were interested in whether the treatment could compel the subjects to take action in real life. We measured the behavior by asking people to donate real currency toward the cause discussed in the article. We entered every survey participant in a raffle to win $100 and asked how much of the raffle winnings the respondent would like to donate toward alleviating air pollution. We stated that we would donate this amount on their behalf and award them the rest, should they win the raffle. This question forced people to put the money where their mouth was, rather than merely proclaiming that they cared. We hypothesized that the treatment group would be more engaged with the topic in the article and donate more toward the cause.

We were not sure what the actual distribution of donation amounts might look like. We tried two different assumptions and the power analysis indicated that either 100 observations would be enough of 700 would be insufficient to achieve the power of 90%.

(Insert assumed dist. and power analyses)

Article Reading Time

For a long article of 1,300 words, compliance was a concern. However, in real life, we wouldn't expect most people to read the whole article. Since the location was indicated in the title of the article, we decided to think of that as the treatment and see how the reading time changes in response.

For the power analysis, we took the average reading time from the pilot, but reduced the standard deviation from the pilot by 20%, taking into account a larger sample size we will collect from the full study. As a result, we calculated that 100 observations is sufficient to achieve a power of 90%.

(insert power analysis graphs here)

Final Methodology

Participant Recruitment and Randomization

Mechanical Turk only allowed us to filter the workers by the state. We recruited 200 survey respondents from Amazon Mechanical Turk in California. In order to filter workers by those living in LA, we instructed the workers to accept the task only if they lived in LA. Additionally, the first question on Qualtrics asked whether the respondent lived in LA and if they didn't, they exited the survey. As a result, all self-identifying LA residents were randomly assigned treatment or control, then given a series of questions.

Outcome Measurement

In conclusion, we measured the following five outcomes. 1. Quiz score 2. Donation amount 3. Reading time 4. Importance rating of issue 5. Credibility rating of article

In order to ensure a high quality response, we have communicated to the survey takers that they please read as they normally would read any other article and they do not cheat on the questions.

Analysis and Results

Exploratory Data Analysis

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.

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

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**)

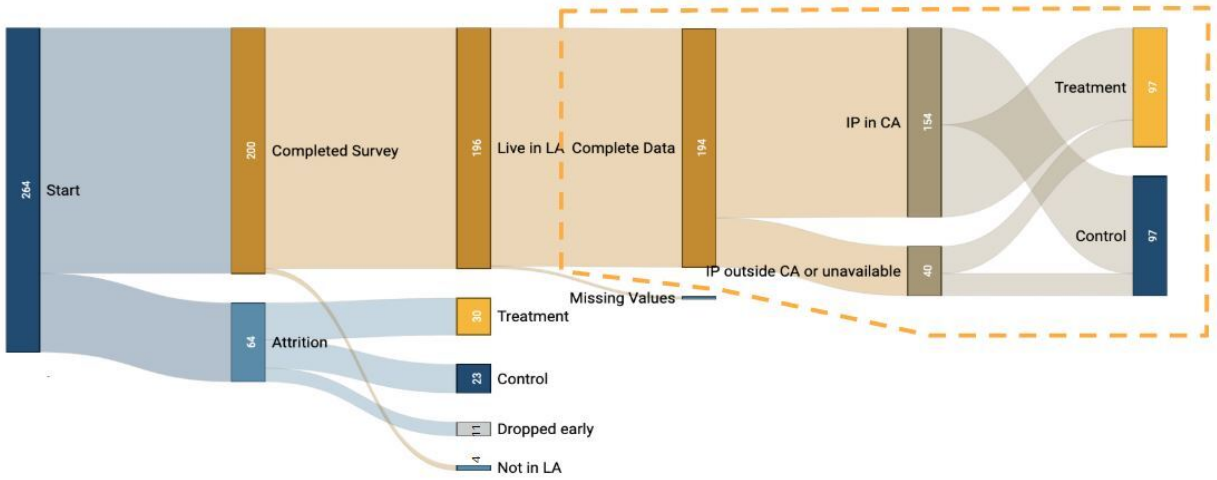
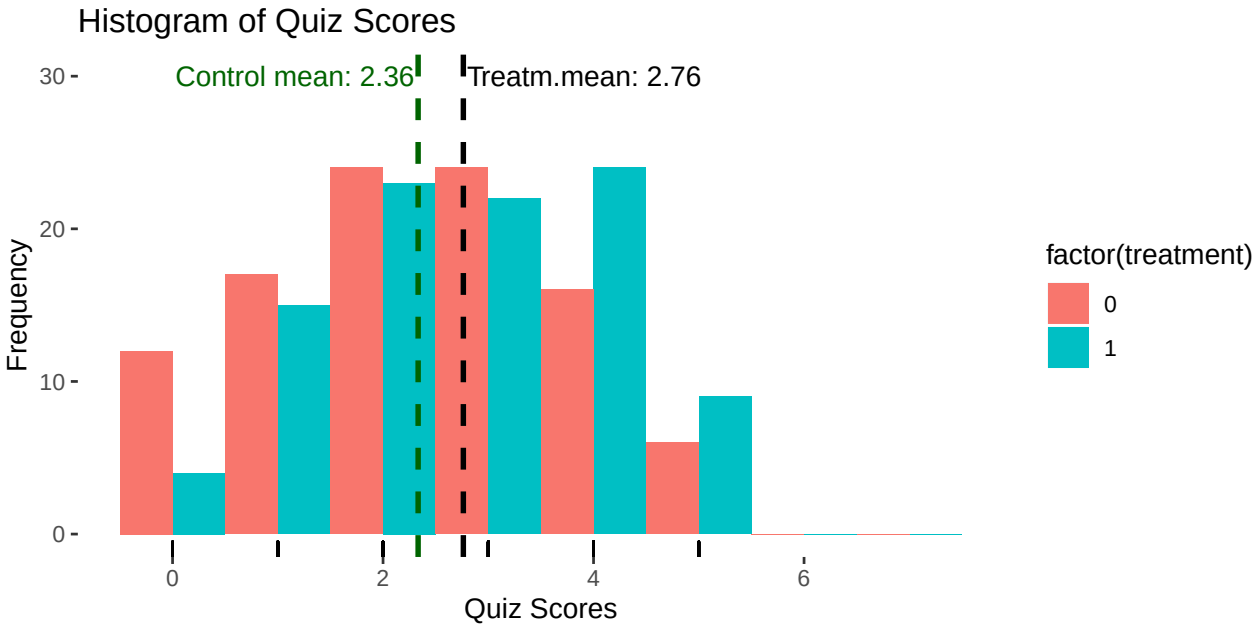


Figure 1: Sankey Diagram of survey data

```
## [1] 2.333333
```

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



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.

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

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

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

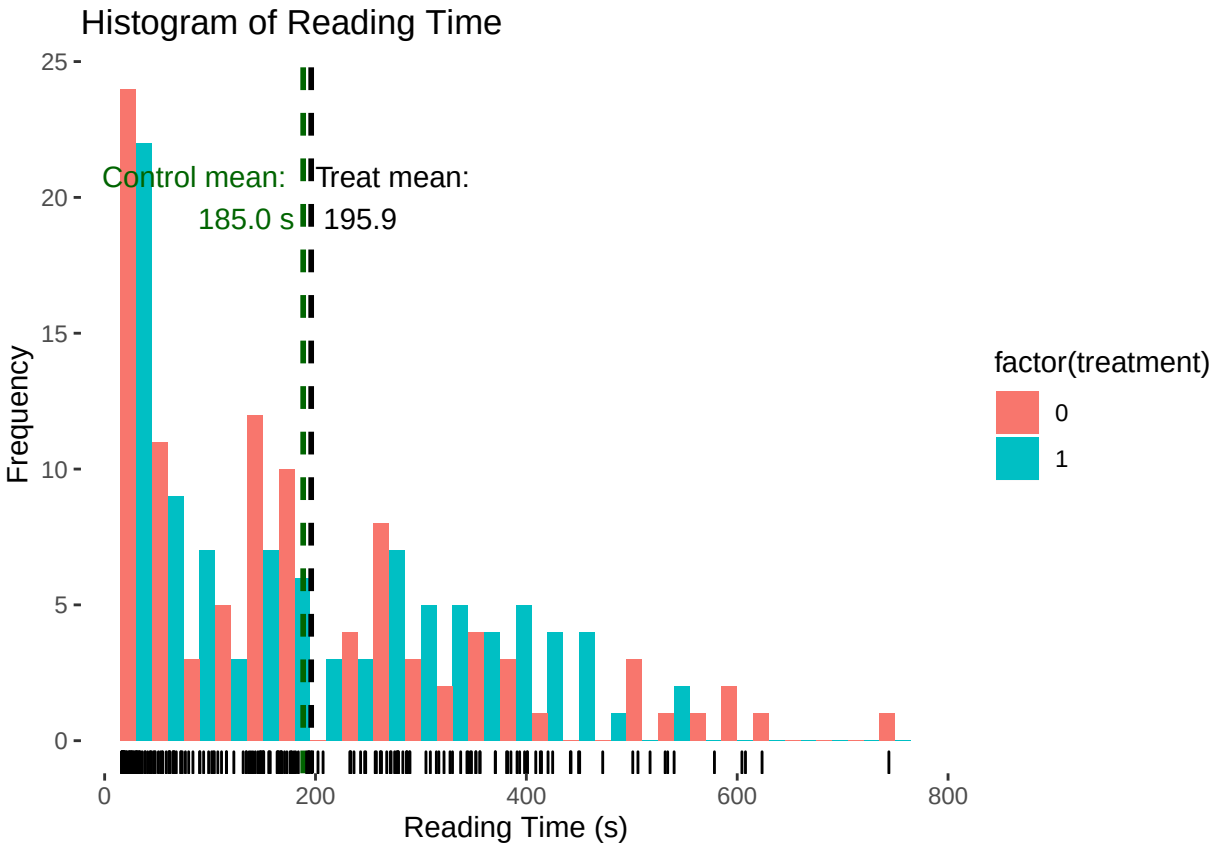
```
## Number of correct answers for question Q_4, is:  34
```

```
## Number of correct answers for question Q_5, is:  110
```

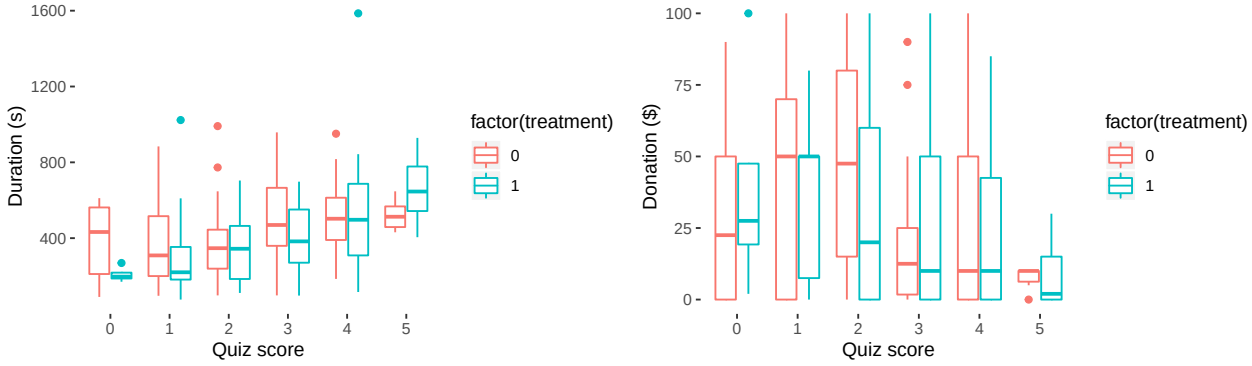
```
## Number of correct answers for question Q_6, is:  119
```

We also check the distribution of reading times for treatment and control. We can see in the graphs below that

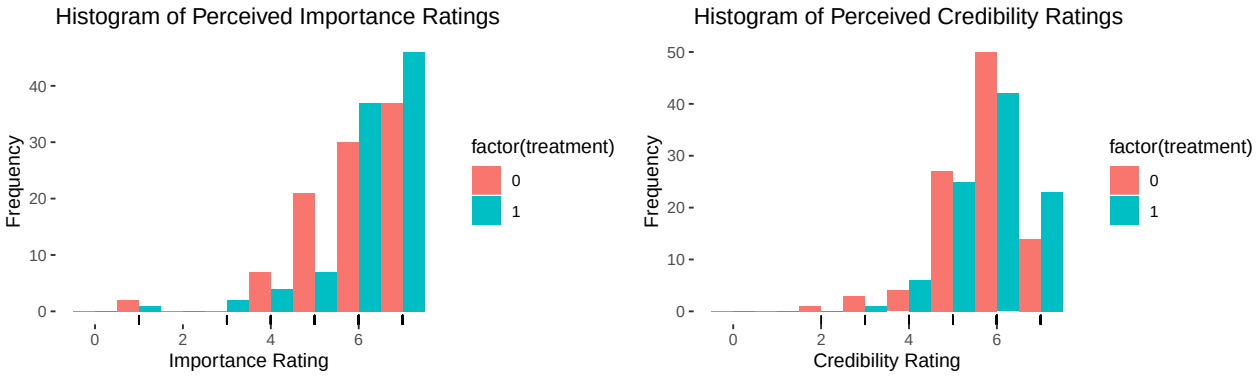
they don't follow a normal distribution. In addition, the actual distribution of donation amounts, seems to have a much smaller difference between treatment and control than either one of our assumed distributions, based on our pilot study (**WE NEED TO ADD GRAPHS OR REFER TO HAERANG'S SECTION HERE**). 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.



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.



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.




```
##
##  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
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d$credibility[d$treatment == 0] and d$credibility[d$treatment == 1]
## W = 4391, p-value = 0.2686
## 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	0.645**	30.599	-3.603
	p = 0.015	p = 0.329	p = 0.510
Observations	113	113	113
R2	0.053	0.009	0.004
Adjusted R2	0.045	-0.0003	-0.005
Residual Std. Error (df = 111)	1.365	164.669	28.734
F Statistic (df = 1; 111)	6.230**	0.963	0.438

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

As shown, we observed a treatment effect of 0.6451 with a p-value of 0.014 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 = 30.5991, p=0.3287) or our Donation ammount outcome variable (ATE = -3.6029, p=0.5093).

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	0.567**
	p = 0.025
Article Read Time (seconds)	0.003***
	p = 0.001
Observations	113
R2	0.144
Adjusted R2	0.128
Residual Std. Error	1.303 (df = 110)
F Statistic	9.252*** (df = 2; 110)

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 0.5665 and is still statistically significant at the 95% conficence level (p = 0.0102).

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.

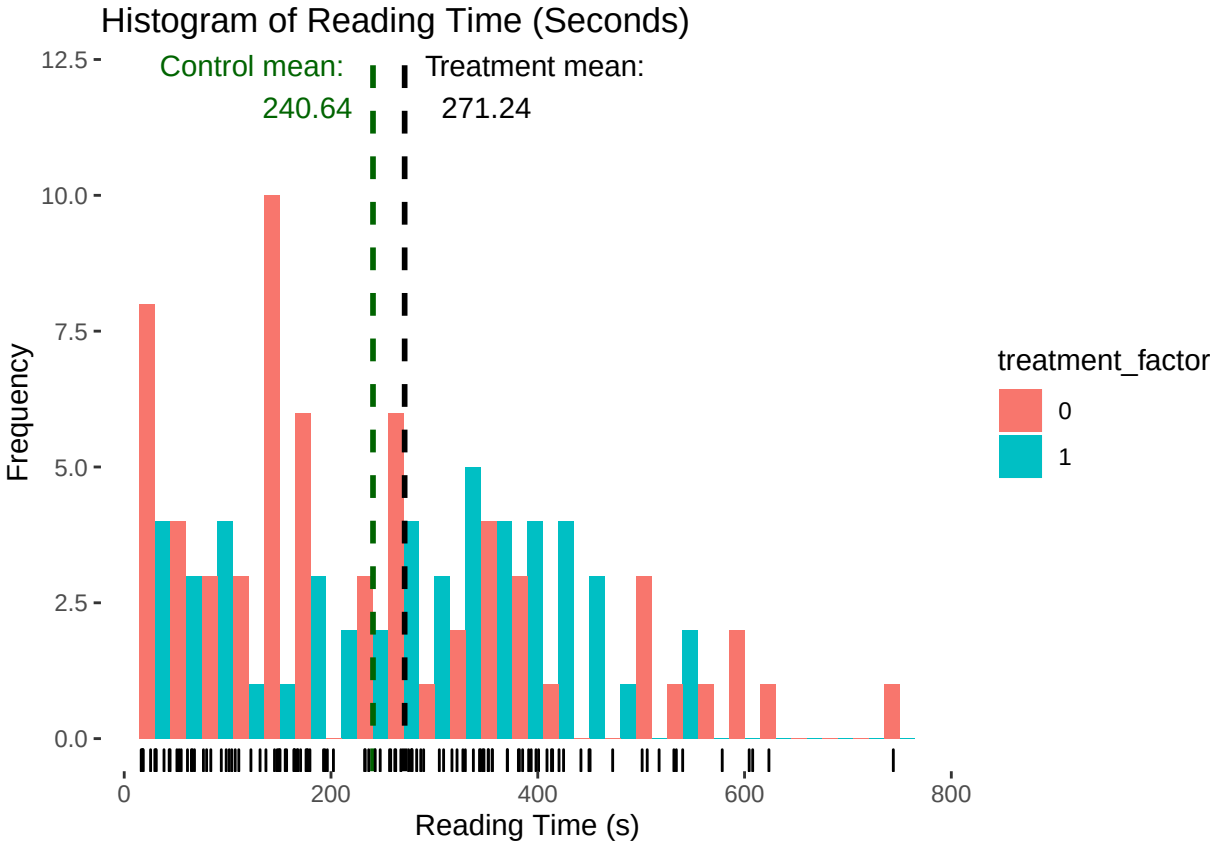
Comparing Treatment Effects

Dependent variable:	Article Read Time (1 Minute bins)	Donation > 0
	(1)	(2)
Treatment	0.667	-0.130
	(0.521)	(0.089)
Observations	113	113
R2	0.015	0.019

Adjusted R2 0.006 0.010
Residual Std. Error (df = 111) 2.752 0.469
F Statistic (df = 1; 111) 1.638 2.147
==========
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.



Generalizability

Science Communication is Broad

Comprehension is Difficult to Quantify

Competing Incentives with Mechanical Turk

Conclusion