Effect of User Reviews on Readers' Perceptions of a Short Story

W241 Final Project Experiment

Jack Workman & Yulia Zamriy

August 3, 2018

1. Introduction

1.1 Research Question

This experiment seeks to answer the question, to what degree, if any, do reviews influence an individual's perception of a short story?

1.2 Motivation

Online reviews have become a strong force in determining business success. If you want to pick a restaurant for dinner, it is quite likely that you will check a few on Google Maps or Yelp to get other people's opinions. If you need to buy something on Amazon and there are multiple options, average score will probably be one of the main factors in your decision. What new movie should you watch tonight? It depends on Rotten Tomatoes score or New York Times review.

However, the power of online reviews does not stop at the decision point. While eating a meal in a 5-star restaurant, will we persuade ourselves that it is worth all five stars even if it is not? After buying a standard plastic storage bin on Amazon with a review score of 3.5 (because somehow perfectly-reviewed bins do not exist), are we going to consider it good-enough but not great despite it being completely functional? If our friends make us watch a poorly reviewed movie, will we find reasons to justify the low score and ignore the good parts of the film?

Are we so reliant on strangers' opinions that it is hard to form our own independent views about consumed products? If we find enough evidence to support this hypothesis, the implications for businesses are considerable. Boosting product reviews would not only drive short-term product sales, but also might help with repeat purchases from "satisfied" consumers.

However, if we find no evidence of consumer compliance to public opinion, it might give us a glimpse of hope that we still can be independent thinkers.

2. Experiment Design

2.1 Hypothesis

The null hypothesis of our experiment is that the average review score of the short story is not relevant to the respondent's review.

The alternative hypothesis is that the average review score of the short story is a significant factor of determining respondent's review (the higher the average review the higher the respondent's score and vice versa).

2.2 Methodology

This experiment was conducted via a Qualtrics online survey. The survey consisted of the following:

- Brief explanation of what to expect
- Short story (about 5-minute read)
- Prompt for user to rate the story
- Questions to assess reader's comprehension of the story
- Question to determine if the responder is familiar with the story

The goal was to, first and foremost, collect the reader's opinion of the story. The additional questions at the end exist to (1) double-check that the user actually read the story and (2) detect any pre-existing bias from the participant. We also added a timer to the short story page to detect any participants who skipped the story.

The survey can be viewed here: https://berkeley.qualtrics.com/jfe/form/SV_5sPNhUpP0zlBssR.

2.3 Treatment

The treatment consisted of a prominently displayed average rating of the story as well as several user reviews gathered from the pilot. The displayed rating took place on the survey page directly before the short story.

The exact rating of the treatment was varied to detect effects in either direction. This resulted in three distinct experimental groups:

1. Control: no average review provided. The users are supposed to rate the story without any external influences.
2. Treatment #1: display high rating (5 out of 6 stars)
3. Treatment #2: display low rating (2 out of 6 stars)

A scale of 6 stars was chosen to avoid giving the participant a middle value. By choosing a scale with an even number of options, the participants are forced to make a conscious decision on whether the story is above average (4 out of 6 stars) or below average (3 out of 6 stars).

2.4 Randomization

Randomization of assigned treatment was implemented as part of the Qualtrics survey. When accessing the survey, Qualtrics randomly assigns each participant to one of the experimental groups and showed him/her the appropriate rating (or no rating if in control). Qualtrics also ensured that participants were equally distributed across groups.

2.5 The Story

The same short story was used for all participants. The story was selected from a science fiction short story website that accepts and publishes short stories. This website has a "random short story" feature that will navigate the user to a random story in its collection. This feature and site were used to ensure that no bias existed in the selection of the story and as a means of selecting a hopefully obscure story.

The story's author gave permission for the use of his story in this experiment.

The story can be viewed here: http://dailysciencefiction.com/hither-and-yon/alternative-history/zachary-morgan-brett/tusks-trunks-and-time-travel.

2.6 Subject Recruitment

Subjects were recruited from personal connections and from Amazon's Mechanical Turk.

3. Experiment Results

3.1 Pilot

We conducted a pilot study prior to launching the first experiment to

1. Test the survey's readiness
2. Gather data to determine the necessary sample size to ensure adequate experimental power

From the pilot's results, we determined that adequate statistical power would come from a sample size of at least size 42. We also made several tweaks to our survey like (1) explicitly requesting that the participants read the entire story and answer all of the questions and (2) easier reading comprehension questions as many participants failed to answer them correctly.

We also confirmed that the story is relatively average. Figure 1 shows a histogram of the ratings collected from the pilot after removing the responses where participants spent less than 60 seconds reading the short story and gave the incorrect response to 2 or more reading comprehension questions.

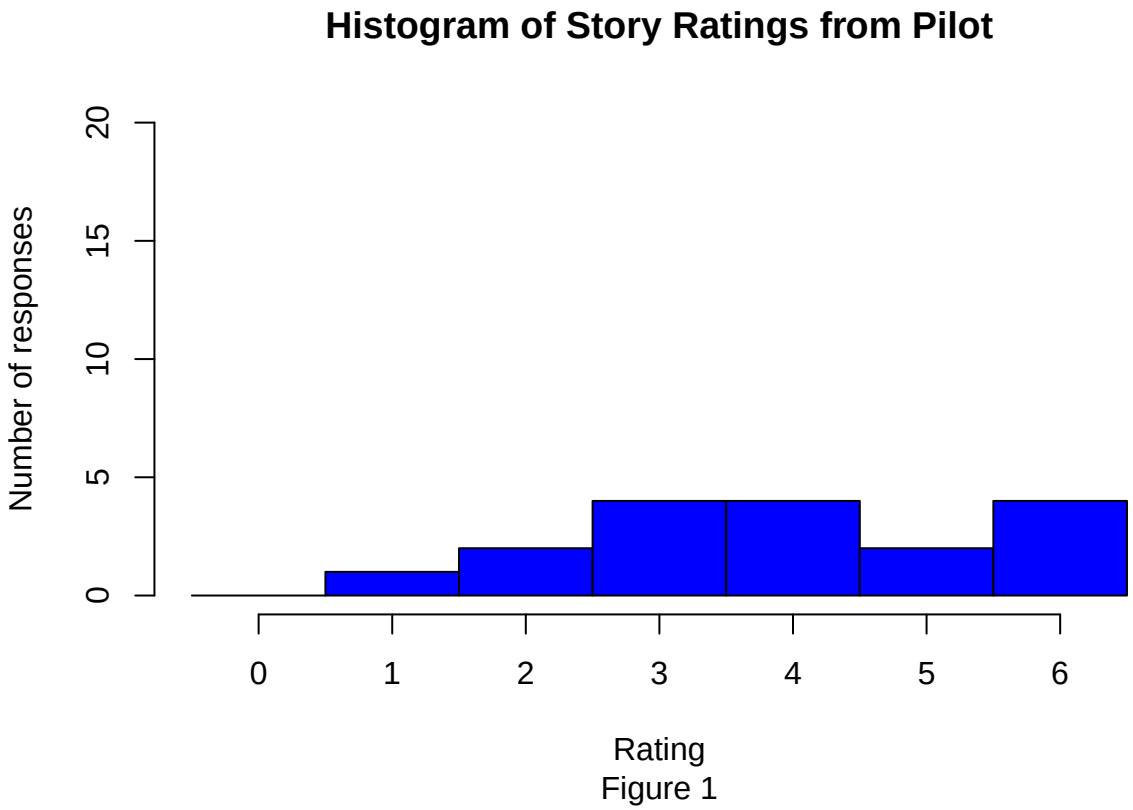


Figure 1

Three rounds of survey requests were issued on Mechanical Turk. The first was for the pilot study and asked for 50 responses. All 50 were given. The second was for the experiment, required a Masters qualification, and asked for 300 responses. Roughly 40 were returned. The third was also for the experiment, did not require a Masters qualification, and asked for 300 responses. All 300 were provided.

3.2. Experiment Data

3.2.1 Data Sources

After conducting the pilot and using the feedback to improve our survey, we launched two additional survey batches on Amazon's Mechanical Turk and sourced participants from personal connections. In total, survey responses were sourced from three different environments:

1. Amazon Mechanical Turk with Masters qualification
2. Amazon Mechanical Turk without Masters qualification
3. Friends & Family (Facebook, LinkedIn, I School Slack)

All survey responses were returned to us via Qualtrics. We were able to distinguish which response belonged to which group thanks to requiring the Mechanical Turk workers to input a code generated by our survey after survey completion. We later joined the datasets together and assigned each participant to their respective group. Table 1 shows the number of responses receveived from each group.

```
##
##  Mturk Masters Mturk Regulars            F&F
##             42            298            179
```

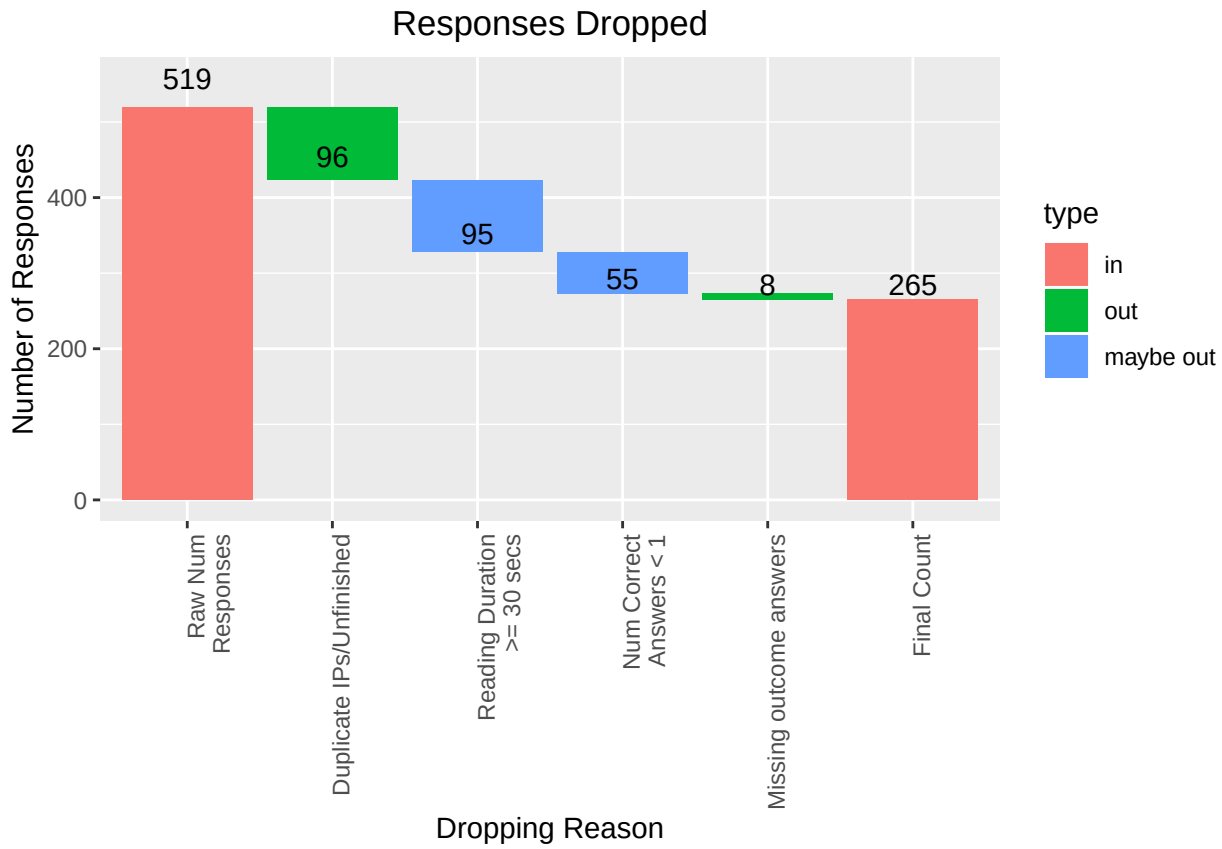
Table 1

3.2.2. Identifying Invalid Responses

Unfortunately, not all participants provided satisfactory responses, and we were forced to label some as invalid. Our criteria for an invalid response is as follows:

- `Status` = 'Spam' or 'Survey Preview'
- Duplicate `IPAddress` occurance - we'll keep the first response for the analysis (alternatively, we can exclude all of them).
- Not finished (progress less than 100%) - These cases need to be investigated for potential bias.
- Time spent reading the story < 60 secs
- Less than 2 reading comprehension questions answered correctly

Figure 2 shows how many survey responses were dropped and for which reasons.



3.2.3 Identifying Treatment and Control Groups

There are three distinct groups in this experiment:

1. Control
2. Treatment - Low Rating
3. Treatment - High Rating

Each particpant's group is determined by Qualtrics at the time the survey is loaded.

```
##
##     Control Treat: High Treat: Low
##   1     147           0          0
##   2       0           0        135
##   3       0         141          0
```

4. Experiment Outcome

We calculated the effect of the treatment with two different methods. The first is with a standard estimated ATE calculation. The results can be seen in Table 3.

The average control rating 4.38. The average low rating was 3.71. The average high rating was 4.26. As you can see, the largest effect was caused by the low rating. Interestingly, the average high rating is lower than the average control suggesting that users are more influenced by lower reviews than high ones.

Group	# of Subjects	% of Total Subjects	Treated Rating	AVG Rating	ATE
Control	95	35.85	na	4.38	0.00
Treatment - Low Rating	86	32.45	2/6 Stars	3.71	-0.67
Treatment - High Rating	84	31.70	5/6 Stars	4.26	-0.12

Table 3

The second calculation method is with linear regression. Linear regression yields an estimated coefficient of -0.1170 for the High Treatment and -0.6696 for the Low Treatment. The Low is highly statistically significant.

```
##
## Call:
## lm(formula = Q4 ~ experiment_group_chr, data = valid_responses[valid_responses$valid ==
##     "Valid", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2619 -0.7093  0.2907  0.7381  2.2907
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.3789     0.1042  42.018  < 2e-16 ***
## experiment_group_chrTreat: High  -0.1170     0.1521  -0.769    0.442
## experiment_group_chrTreat: Low   -0.6696     0.1512  -4.429 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 262 degrees of freedom
## Multiple R-squared:  0.07703,    Adjusted R-squared:  0.06999
## F-statistic: 10.93 on 2 and 262 DF,  p-value: 2.751e-05
```

```
##                                      2.5 %     97.5 %
## (Intercept)                      4.1737403  4.5841545
## experiment_group_chrTreat: High -0.4165995  0.1825143
## experiment_group_chrTreat: Low  -0.9673475 -0.3719426
```

5. Discussion

5.1 Potential Experimental Pifalls

5.2 Generalizability of Results

5.3 Mediation of Results

6. Conclusion