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

W241 Final Project Experiment

Jack Workman & Yulia Zamriy

August 13, 2018

Contents

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 (on 1 to 6 scale)
- 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 responder actually read the story and (2) detect any pre-existing bias from the participant (if they have read the story before). We also added a timer to the short story page to detect any participants who skipped the story.

The survey can be viewed here.

2.3 Treatment

The treatment consisted of a prominently displayed average rating of the story as well as several reader 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 readers 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).

The 5/6 and 2/6 ratings in the treatment groups have been chosen for their perceived credibility. Even though they seem unbalanced, they are both one star away from the extremes (it is very rare to see 0 stars, usually the lowest rating is 1 out of the max available).

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 shows him/her the appropriate rating (or no rating if in control). Qualtrics also promises that participants will be 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.

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 our main experiment. It covered only the control version of the survey and was designed to:

1. Test the survey's readiness
2. Gather data to determine the necessary sample size to ensure adequate experimental power
3. Determine average rating that would be appropriate for treatment groups

From the pilot's results, we determined that adequate statistical power would come from a sample size of at least size 42. The following parameters were used for this calculation:

- Desired effect size: 1 (difference in average score between treatment and control groups)
- Standard deviation in the pilot study: 1.56
- Alpha: 0.05
- Power: 0.9

```r
# Built-in formula
d = ate/sigma # assuming sigma is pooled standard deviation (equal variance)
pwr.t.test(d = d,
           sig.level = alpha,
           power = 1-beta,
           type = 'two.sample',
           alternative = 'greater')
```

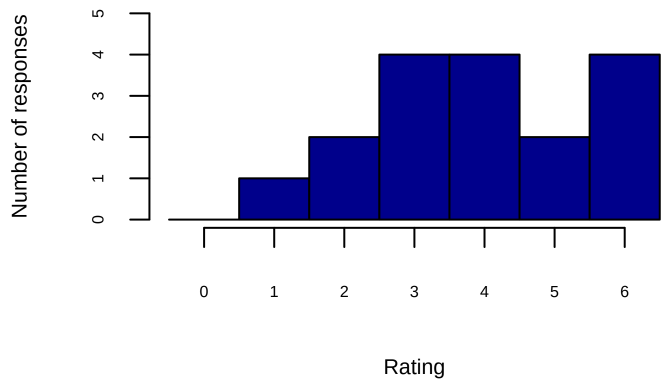
```
##
##      Two-sample t test power calculation
##
##               n = 42.37973
##               d = 0.6409962
##       sig.level = 0.05
##           power = 0.9
##     alternative = greater
##
## NOTE: n is number in *each* group
```

Based on pilot outcome, we 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 did not satisfy our valid response criteria:

- spent less than 60 seconds reading the short story and
- gave the incorrect response to 2 or more reading comprehension questions

Figure 1. Histogram of Story Ratings from Pilot



3.2. Experiment Data

3.2.1 Data Sources

For our main experiment, survey responses were sourced from three different environments:

1. Amazon Mechanical Turk with Masters qualification (25 cents per response)

2. Amazon Mechanical Turk without Masters qualification (25 cents per response)
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.

Table 1. Responses by Source Group

Source	Count
Mturk Masters	42
Mturk Regulars	298
F&F	179

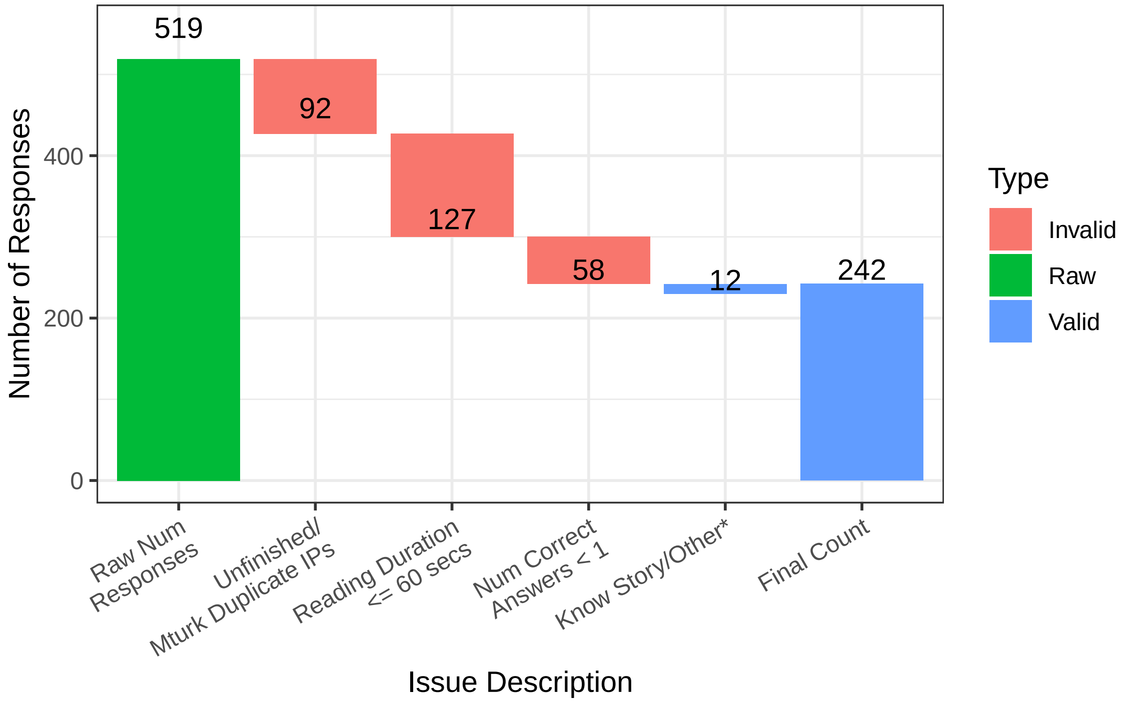
3.2.2. Identifying Invalid Responses

Unfortunately, about half of the 519 responses had to be flagged as invalid. Our criteria for an invalid response is as follows:

- `Status` = 'Spam' or 'Survey Preview'
- Not finished (progress less than 100%). We will remove these from the analysis because they contain no responses to any questions
- Duplicate `IPAddress` occurance. We drop duplicate responses among `Mturk` responders (high likelihood of fraud), but keep them for 'Friends and Family' as members of the same household could use the same computer. However, this potentially could result in a spill-over effect
- Time spent reading the story < 60 secs. This experiment relies entirely on the assumption that the subjects read the short story. To ensure this, we added a hidden timer to track how long each participant spends on the short story page itself. The short story is 990 words long, so any subject with less than 60 seconds, a reading speed of 990, wpm will be dropped. Given that the adult average reading speed is about 200 wpm, we believe that this is more than justified.
- No reading comprehension questions answered correctly. The survey contains three reading comprehension questions to test the reader's understanding of the story. These questions are designed to be extremely basic and high-level. In fact, the questions were made easier after the pilot as those were deemed to be too difficult. If the subject read the story, then they should be able to answer these questions. Since no one is perfect, we are electing to include in the analysis only the subjects that answered at least 1 question correctly.

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

Figure 2. Validity of Survey Responses



3.2.3. Balance Checks for Invalid Responses

There are three distinct groups in this experiment:

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

In this section, we will check if responses marked as invalid were fairly distributed across the above three groups.

Qualtrics allocated surverys at random order to the above three groups. As we can see from `Table 2` there was some slight inbalance as `Treat: Low` received a bit lower than fair share of responses. We'll assume it's due to random noise.

Table 2. Qualtrics Treatment Group Assignment

	Control	Treat: High	Treat: Low
Responses	178.00	179.00	162.00
Percent	0.34	0.34	0.31

As indicated in `Figure 2`, we dropped 92 responses because they either did not finish the entire survey or did not answer the main question. `Table 3` checks the distribution of these responses across treatment groups. The only issue that stands out is that an unproportionally low number of unfinished surveys were in the `Treat: Low` group, but it is consistent with the fact that Qualtrics allocation was not even.

Table 3. Unfinished Surveys Balance Check

	Control	Treat: High	Treat: Low
Missing outcome	4	4	6
Not Finished	25	28	17
Duplicate IP	3	1	4
Valid	146	146	135

After we drop the 92 responses, we exclude 185 cases that either spent less than 60 seconds on the survey or did not answer at least 1 reading comprehension question correctly. The distribution of responses across groups is consistent with the original allocation by Qualtrics (`Table 4`)

Table 4. Valid Responses Balance Check

	Control	Treat: High	Treat: Low
Undertime/Failed RC	60.00	64.00	61.00
Valid	86.00	82.00	74.00
Percent	0.36	0.34	0.31

The last check is to see how valid responses were distributed across source groups (`Table 5`). Interestingly enough, `Treat: Low` group has more responses in the `Friends and Family` group compared to `Treat: High` and `Control`. We do not have a strong hypothesis to explain this inconsistency and we will assume that it's random.

Table 5. Source Balance Check

	Control	Treat: High	Treat: Low
Mturk Masters	13	10	9
Mturk Regulars	48	47	35
F&F	25	25	30

4. Experiment Outcome

Our final analysis sample consists of 242 valid responses. However, we will check if including invalid responses impacts the results.

4.1 Outcome Distribution

Before calculating ATE as a difference in average rating across treatment groups, we will check the distribution of this variable.

Responders had to rate the story on a scale from 1 to 6. We decided to use an even scale to force responders pick a side instead of providing a neutral response. Based on `Figure 3a` for `Control` group, average rating without any treatment was above 3.5 (expected average). This indicates that the story was slightly better than expected

`Treat High` group had the highest proportion of 5-star ratings, but it was not enough to increase the group's average rating compared to the `Control` group. `Treat Low` group, on the other hand, had a lower average rating compared to the other two groups.

Figure 3a. Control: Rating Distribution

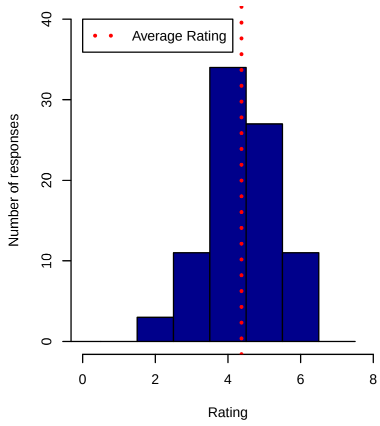


Figure 3b. Treat Low: Rating Distribution

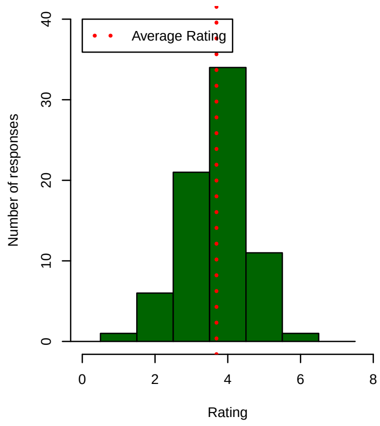
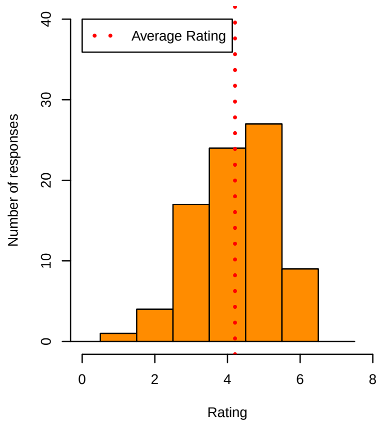


Figure 3c. Treat High: Rating Distribution



4.2 Average Treatment Effect

4.2.1 Manual Calculation

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 6`.

The average control rating was 4.37. The average low rating was 3.69. The average high rating was 4.21. As you can see, the largest effect (in absolute terms) was caused by the low rating treatment. Interestingly, the average high rating is slightly lower than the average control. These results suggest that users are significantly more influenced by lower reviews than high ones.

Table 6. Average Treatment Effect

Group	# of Subjects	% of Total Subjects	Treated Rating	AVG Rating	ATE
Control	86	35.54	NA	4.37	
Treatment - Low Rating	74	30.58	2/6 Stars	3.69	-0.68
Treatment - High Rating	82	33.88	5/6 Stars	4.21	-0.16

4.2.2 Regression Analysis

The second calculation method is with linear regression. Linear regression yields (`Table 7`) an estimated coefficient of -0.16 (robust std.error: 0.16) for the `Treat High` and -0.68 (robust std.error: 0.15) for the `Treat Low`. The Low group coefficient is highly statistically significant with a confidence interval of (-0.98, -0.39).

```
##
## Table 7. Regression Estimated ATE
## =================================================
##                          Dependent variable:
##                     -----------------------------
##                            Average Rating
##                         Valid Responses Only
## -------------------------------------------------
## Treat: High                    -0.165
##                                (0.164)
## Treat: Low                    -0.683***
##                                (0.152)
## Constant                      4.372***
##                                (0.107)
## -------------------------------------------------
## Observations                     242
## Residual Std. Error       1.013 (df = 239)
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

4.2.3 Regression Analysis and Invalid Responses

Since we excluded a fair share of responses due to serious validity issues, we want to check if including them changes our results. `Table 8` compared outputs of 3 models:

1. Our main regression on 242 valid responses (same as `Table 7`)
2. Same model specification but with all 427 responses included
3. Model with all 427 responses and additional controls for invalid responses

```
##
## Table 8. Regression Estimated ATE
## ===================================================================================
##                                               Dependent variable:
##                              ------------------------------------------------------
##                                                  Average Rating
##                              Valid Responses Only  All Responses    All Responses
##                                      (1)                (2)              (3)
## -----------------------------------------------------------------------------------
## Treat: High                         -0.165            -0.041           -0.039
##                                    (0.164)           (0.130)          (0.131)
## Treat: Low                         -0.683***         -0.375**         -0.411**
##                                    (0.152)           (0.129)          (0.127)
## Finished in < 60 sec                                                   0.116
##                                                                       (0.120)
## Failed Reading Comprehension                                          -0.572**
##                                                                       (0.186)
## Constant                           4.372***          4.212***         4.253***
##                                    (0.107)           (0.088)          (0.097)
## -----------------------------------------------------------------------------------
## Observations                         242               427              427
## Residual Std. Error            1.013 (df = 239)   1.103 (df = 424) 1.090 (df = 422)
## ===================================================================================
## Note:                                                 *p<0.05; **p<0.01; ***p<0.001
```

Unfortunately, the results in `Table 8` indicate that:

- After including invalid responses the `Treat Low` effect is around 50% lower, but still statistically significant.
- Responders that failed all 3 reading comprehension questions had significantly lower average rating compared to other responders with coefficient -0.57 (0.19).

However, if we check `Table 9` below, the majority of responders that failed reading comprehension were from `MTurk Regulars` group. Therefore, we can suggest two hypothesis for the above results:

1. `MTurk Regulars` did not read the story carefully and, hence, did not have strong opinion on it. They then picked the score that was closer to the expected average rating
2. `MTurk Regulars` did not read the story at all (potentially bots) and picked the score that was closer to the expected average rating

Regardless of the explanations, the results are still consistent: the impact of Low ratings are statistically significant, while High ratings do not make people change their minds.

Table 9. Number of Responders That Failed All Reading Comprehension Questions

	Response Count			**Average Rating**		
Source	Control	Treat: High	Treat: Low	Control	Treat: High	Treat: Low
Mturk Masters	2	NA	NA	3.00	NA	NA
Mturk Regulars	18	20	12	4.06	3.65	2.92
F&F	2	3	1	4.50	4.67	3.00

4.2.4 Source Group Results

Another interesting hypothesis that we can test is whether results differ by source group (`Table 10`).

```
##
## Table 10. Regression Estimated ATE by Source Group
## ====================================================================
##                                   Dependent variable:
##                     ------------------------------------------------
##                                      Average Rating
##                      Mturk Masters   Mturk Regulars        F&F
##                           (1)             (2)              (3)
## --------------------------------------------------------------------
## Treat: High              0.300           -0.285           -0.160
##                         (0.497)         (0.213)          (0.299)
## Treat: Low              -0.778          -1.004***         -0.187
##                         (0.464)         (0.226)          (0.231)
## Constant               4.000***         4.604***         4.120***
##                         (0.312)         (0.134)          (0.198)
## --------------------------------------------------------------------
## Observations              32              130              80
## Residual Std. Error 1.077 (df = 29) 1.035 (df = 127) 0.908 (df = 77)
## ====================================================================
## Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

These results suggest that:

- In none of the three groups, the `High` rating impacted people's opinions about the story (as compared to the `Control` group rating)
- In `MTurk` groups, the `Low` rating was 0.8 to 1 lower than the average `Control` group score. However, the effect estimate for the `Masters` group is not statistically significant mostly due to low sample size

- For the F&F group, neither of the treatments made a difference. Their average ratings were statstically indistinguishable across all treatment and control groups

It is hard to explain these results without knowing MTurk responders audience. However, we can speculate with high confidence that Friends & Family audience has higher education and employment levels than people who earn money on MTurk. That might influence how they read and perceive the story. And it might be that they are much better at forming and staying with their opinions.

5. Conclusion

After conducting a survey with the aim of measuring the impact of high and low numerical ratings on a reader's perception of a short story, we conclude that the evidence suggests readers are, in fact, influenced by negative reviews as statistical analysis of both a manual calculation and regression analysis of the treatment effect yielded a highly statistically significant coefficient of -0.68 stars (on a scale of 6 stars) for readers shown a low numerical rating prior to reading the short story. Positive reviews, on the other hand, appear to not be influential as no statistically significant relationship was found.

There are several concerns worth noting alongside these results. First, we have some doubt on whether all participants actually read the story. To catch this potential escape, we built several controls into our experimental design including a reading timer and reading comprehension questions. Based on the data gathered from these controls, we invalidated about half of our responses - a regretfully large amount. To be sure that this did not negatively influence the accuracy of our results, we ran an additional regression on the full dataset comprised of invalid and valid responses (Section 4.2.3). The significance of the results, albeit with a smaller treatment effect, proved consistent. The second concern is uncertainty of the representativeness of the sample. The majority of our participants came from Amazon's Mechanical Turk, and we know little of their demographic and educational backgrounds. In Section 4.2.4, we explored the potential differences between our participant sources with a regression on each source's set of responses. The MTurk Masters source group sample size was too small to make any conclusions, the MTurk Regulars source group yielded the same observed significance level for negative reviews as the original analysis, and Friends & Family showed no treatment effect at all. This is potentially because Friends & Family were aware of the possibility of being treated while taking the survey. Ultimately, more experimentation is required to reach any further conclusions on the differences between these experimental groups.

How do the results of this experiment generalize to other industries and domains? In today's society, ratings and reviews seem to be everywhere. Most prominently, they can be found in the commerical sector on sites such as Amazon, Google, or Yelp. While a short story is not directly comparable to a movie or restaurant experience, we believe that the pyschological impact of seeing a review prior to the targeted event persists across these experiences.

From a business perspective, the takeaway of this experiment is clear: avoid negative reviews. This is not particularly shocking as many might consider it common sense that a positive reputation is likely to lead to more customers. From an individual perspective, this experiment suggests that we can be unknowingly biased by negative reviews which is something to keep in mind while perusing Amazon for a new product or searching Yelp for your next meal.