Are More Reviews Better? Evidence from a Policy Change on the Steam Store

Adam Di Lizia

November 2022

1 Background

Steam is a popular platform for PC users to download video games over the internet. By 2021 it had over 130 million active monthly users (Wikipedia) and 75 Percent of the global PC game market. It is not just purely a marketplace: similar to sites like Amazon, users who have purchased a game on Steam can leave a review. They can either recommend or not recommend the game to others, and they can write as much text as they want as a justification of their review.

Figure 1: An example of a simple review of the game 'Grand Theft Auto 5." Displayed is the user, the playtime at time of review, time of review, number of games the user has in their account, how helpful or funny the review is and number of other reviews they have left.

On October 30 2019 the 'Library Update" is rolled out globally across Steam. Prior to the update, to leave a review on a game you had to navigate to the store page of the game you owned, scroll down to the review section and then write the review. After this update, Steam added a feature where exiting a game that

you have yet to review triggers a pop up asking you to leave a review. This pop up allows the user to fully write and publish a review without any navigation on their part, and is much quicker to do. This had a drastic increase in the number of reviews left by users. In Steam's own words:
"The updated library also introduced some new ways to prompt the user to review or re-review a game they have been playing based on certain criteria, which substantially increased the number of reviews written. We previously saw around 17 thousand new reviews posted per day across all games, but since the Library update, we now see 70 thousand reviews per day, for an increase of over 300 Percent."

This research project hopes to leverage the policy change to see whether lowering the barrier to entry for reviews on a platform has consequences on the types of people who review, the average rating and the length and quality of reviews.

2 Motivation

Ideally, a consumer faced with thousands of reviews should be able to make an informed decision. However, as people must select into providing a review, often only the more extreme experiences are represented. Someone with a very bad or very good experience is more likely to leave a review than someone with a mediocre one. Therefore, the policy change on Steam represents a direct opportunity to measure what happens when the cost of submitting a review is lowered. There is still self selection, but much less so than before. It would therefore be an interesting test of the prevailing theory of extreme reviews, and furthermore allow an analysis of how useful lowering the cost of entry might be to a platform. While there is research on platform reviews, and on the general extreme distribution of reviews across platforms, this project is in my knowledge the first to capture a change on the same platform. This could provide new evidence on how a platform could work to reduce the bias in its reviews.

In addition, Steam as a platform has been largely ignored by the economics profession. To my knowledge there is not a single economics paper harnessing the incredibly rich review data available on the platform. Therefore this data can provide new evidence and analyses on how platforms might better shape reviews for their own or their consumers' benefit.

3 Research Question

The broad question this project aims to answer is one about the changing of reviewer types. Specifically, as the cost of leaving a review is lowered, do we see a new type of reviewer? The literature would suggest this new type has less extreme opinions as before, and perhaps we might hypothesise that they

are more likely to leave a "true" review (positive if the game is largely positive, negative if the game is largely negative). We might also expect this person to leave shorter, less useful reviews compared to those who spent the time to find the game in the store manually and leave a review. Finally, do we see more people leaving reviews after the policy change who have never reviewed before? This would lend credence to the idea that a new type of reviewer has been coaxed into the review space.

Specifically, this project asks how do reviews on the steam store change post-policy. How does lowering the cost of leaving a review affect the pool of potential reviewers? And does the average review get more positive or truthful? And what happens to the written reviews?

There seems to definitely be scope for a theoretical model to motivate an exact mechanism here. A model where reviewers care about how accurate steam reviews are might lead those with more extreme opinions to bear the cost of leaving a review as there is more scope for Bayesian updating with respect to an average prior, whereas someone with an average view feels they have little to add. Equally, it could be a simple emotional issue: people who really hate/enjoy a game want to share that more. Seeing if reviews are more truthful or if they have a positive/negative emotional slant might also help us understand the mechanism behind extreme reviews.

4 Papers

There is no research I could find that does exactly what I propose for a different review site. However, there are some useful papers that reflect different parts of the project well.

The first of these is: "The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications" (February 15, 2019 Columbia Business School). This is an incredibly useful paper which not only surveys most of the literature on extreme reviewing up to that point but performs a comprehensive analysis on the general extremeness and polarity of reviews on over 20 different sites (no Steam though). They find strong evidence of self selection in online reviews but show that the level of polarity of reviews varies among different sites. This paper is useful as not only does it do an excellent job of motivating my mechanism and research, but also is a good handbook of working with any online review data. For example, they find that the number of reviews per reviewer is a good proxy for level of self selection. This is data I have, but it is useful to know what should be controlled for and broadly what I might expect to see from different variables.

The next paper is "An empirical study of game reviews on the Steam platform" (Empirical Software Engineering 2019). It is not an economics paper, however it is useful as it is the only piece of research I can find on Steam reviews and does an incredibly thorough job of statistically describing steam reviews and describing in detail their scraping process and all the data they obtained. It is useful as they describe a process for calculating the readability of a review, which is something that might change post policy. They also show that the length of positive reviews and negative reviews are equal. Perhaps this might also change post review. It finds a whole host of descriptive statistics and distributions on what reviews are about, play time before leaving a review etc. that can be calculated on my longer data set (theirs went only to 2016). This would allow me to see whether these statistics significantly change post policy and ensure I am able to measure as much as I can.

The final paper of use is: "Estimating dynamic treatment effects in event studies with heterogeneous treatment" (Journal of Econometrics 2021). This is important to ensure I have a robust and consistent estimating procedure. One potential method to secure identification involves a staggered treatment of another variable which requires careful estimation. In general this paper will be a good guideline for my estimation and inference.

5 Empirical Strategy

A crucial feature of my data is that games have review "cycles" which are all out of sync as different games were released at different times. Therefore the policy might have occurred only a week after one game was released and 3 years after another game was released. This allows me to control for the review cycle fixed effects (e.g do games get a bump or drop in reviews or their characteristics after some time).

However, as the policy was universal it is essentially a time fixed effect. It is important to allay fears that this is just a shock to reviews unrelated to the policy change. Since I have reviews up to the minute of posting, and lots of reviews across many games, I can limit the post review analysis period from a week to n weeks after. This would reduce the chance any other shocks far into the future pollute my results. Furthermore, I can run placebo tests on other time fixed effects after controlling for seasonality, time trends and game sales to ensure that the policy effect is significantly larger and that I am not just picking up a shock to reviews. This is likely to be identified as you can see in the data below:

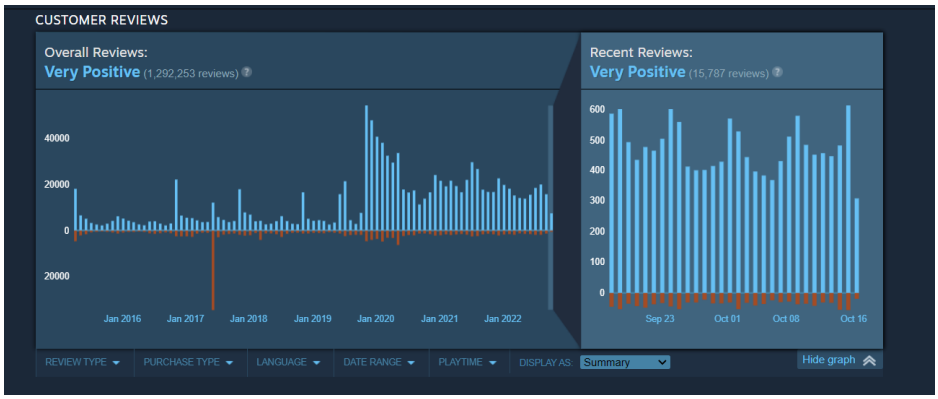

Figure 2: The basic trend of number of positive and negative reviews for Grand Theft Auto 5 available on Steam. The large spike occurs after the policy change and is significantly larger than any previous spike except for a large negative one due to a 'review bomb" (explained in the data section)

The regression I plan to run will be a fixed effects regression:

$$y_{it} = a_i + rc_\tau + \beta POST_{it} + \delta REDUC_{it} + z_{it}'\gamma + \epsilon_{it}$$

Where i is an observed game review and t is the time period to be aggregated upon. y_{it} is my outcome variable. rc_τ are review cycle fixed effects and a_i are game fixed effects. $POST_{it}$ is 1 after the policy is implemented. z_{it} is a vector of controls including sale price, displayed average review and other characteristics of parent game such as patches, number of other reviews by that user, playtime at time of review, number of games in account and other characteristics of the reviewer from their steam profile.

$REDUC_{it}$ is the percentage price reduction at time t (different to absolute price) i.e is there a sale on for game i. I mention this separately as since sales occur at different times and review cycles across my sample, a careful diff-in-diff will allow me to see the causality of sales on my outcome variables. It is possible that a sale also has some affect on reviewer types as the cost of the product falling for a brief time might attract a different type of gamer. Therefore it will certainly be worth analysing too. More importantly, if there are enough sales around the window of my policy change, then I could use these to control for my time fixed effects and fully identify the policy effect.

The beta above will not be a treatment effect but rather the weighted average treatment effect across each game in my sample. This potentially motivates splitting this effect up by different categories e.g large games vs smaller indie games. This allows for heterogeneity of effects to be accounted for by running the regression of interest on these different sub-populations.

My outcome here will be multiple different statistics. First, whether reviews are positive or negative. Review truth/accuracy can be measured by seeing how close the Steam reviews match with reviews from metacritic, a website which aggregates all independent journalistic reviews of games. I will also regress the length and reading age of the review (calculated as in Lin et al), the number of reviews (known to increase but a good benchmark), the user submitted helpfulness of the review and the ratio of any of these characteristics between positive and negative reviews. This is important as it could be that the average does not change but the discrepancy between positive and negative reviews does. There is further scope for analysis of the review text using machine learning methods. One potential example is using review sentiment (how positive or negative the review is written) rather than review score as the sentiment being continuous might pick up more nuanced changes after the policy.

As stated earlier, I will run placebo tests on different non-policy dates to ensure my method does not just pick up noise. I will also have to potentially modify the regression slightly depending on the outcome variable e.g review score being 1 or 0 motivates a probit model.

6 Data

Data collection is one of the most important parts of this project. I plan on web-scraping the Steam review page for every game that existed before the policy change. I will start with a couple big games such as Grand Theft Auto 5, and slowly expand depending on the success of initial results. Thankfully, the code to scrape Steam is freely available online from machine learning students who train classifiers on Steam reviews. Therefore, the next step is for me to run this code for a single game, and eventually more. I will start with Grand Theft Auto 5 as it is the largest grossing media property of all time and has a large number of reviews weekly (thousands), and then expand to other best selling and popular games. Data will be cleaned and saved in a csv file with each review and its characteristics. Data on games and individuals specifically can easily be merged with python giving the full data set required for regression.

Steam DB is a database online that has the history of prices for all steam games and all patch note history. This is important as these are things I cannot directly scrape as easily but this data is already available as csv files to be merged with my other data across time. Metacritic can easily be web-scraped as it simply shows the score out of 100 that critics give the game on average.

Once this is done however I will have an incredibly large and rich data set of every steam review, the user who posted them and characteristics of the game and poster. Thanks to Steam's built in positive-to-negative review chart (pictured above) I can clearly see there will be a large variation in the games in

my sample. I also have incredibly fine grain time variation as the exact time of the review is included on Steam. I cannot say before analysis how much length or helpfulness changes, but I know at least within a snapshot of time there is a large variation between length and review helpfulness. Some reviews are pages long with 100s of people finding them helpful and others are one word that no one finds helpful.

I have been using Steam for many years, so there are important issues with the data that are not immediately obvious. One main one is that of early-access games. This is a type of game that is released early on Steam before it is fully finished, with the developer promising to finish the game and respond to user feedback. As a result reviews for this game will change as the games quality does, and the reviewer expects their review to have an actual impact on the games quality potentially as the developer might read and implement a change. Therefore these ought to be dropped from the data set.

Similarly, Another thing to consider are patches to a game. This refers to the industry practice of releasing a free update to a game post-launch that changes some aspect of the game. The word "patch" is related to the fact that these usually are in the forms of minor bug fixes. Before around 2015, this would have had zero effect on the reviews to a game but as the industry tends towards a live service style model, many games are released in incredibly unfinished states either due to glitches or low amounts of content. Therefore a patch could include free new content to a game or make it much more playable. Both of these could affect the review score. Therefore, taking patch history from Steam DB I can ensure that there are no important patches to any games in my sample around the policy change (otherwise they must be dropped). In addition I can control for patches across the rest of my sample.

Another issue is review-bombing. This is when a company or game does something that a wide proportion of the consumer base does not like and as a result there is a call to action on social media to leave negative reviews as a form of punishment. This could be in response to the game itself having features removed or a grievance toward the parent company or publisher. For example, the large negative spike in Fig 2 is a review bomb in response to publisher Take Two attempting to limit mod support for the game. These review bombs need to be removed from the data set as they create a different type of review for a different purpose. They can be identified both by a large spike in number of reviews and they should only exist on one game. Note they can also be positive too. When a popular YouTuber makes a video on a game years after release this can flood the game with positive reviews that thank the YouTuber for sending them there. I also want to identify positive review-bombs that are ridiculously high via the same criteria and remove them too.

My ideal data set would be one where the library update only affected some games and not others to also control for time fixed effects. However, due to the large amount of controls and the general nature of reviews, my hope is that after controlling for seasonality, sales and trends too there should be no time fixed effects of a similar magnitude. This is a theoretical prediction in that I do not see what else could affect *every games* overall review characteristics at the same time post controls other than some other policy change. However, if I do still see time fixed effects of a similar magnitude to my policy effects then clearly identification of the effect is unlikely. A potential solution is if I can leverage the differences in sale times to back out time fixed effects which would give me very clean identification. This requires enough sales around my period of interest however.

Another change to the data that would be ideal would be if reviews were out of 5 rather than 0 or 1. This would much more easily allow me to see a reduction in variance or a slight increase for the average person. Together with a model and other outcomes this can be hopefully mitigated. The best I can do is likely a continuous sentiment score based on the text and the review posted.

7 What Could Go Wrong?

There are many potentially pitfalls with my proposal which I would like to mitigate. Crucially, the main issue is in data collection. The code is prepared, but the resources required to scrape every game will be intense. Therefore it is vital I start with one or two big games to ensure I do not waste time if the effect is relatively negligible after controls. It seems unlikely that I will not see any effect based off the graphs available to me, but it could also be that this effect is equal in size to others and that my placebo tests fail.

One area I may get null results however is in other outcome variables such as length or helpfulness of reviews. This is a problem as since I only see a 1 or 0 for review score so for any mechanism of a different "type" of reviewer to be shown I would expect to see some change in something else. However, I definitely do expect some null results. If i see that the ratio between review length of positive and negative reviews is unchanged that is not particularly a problem. Nonetheless it definitely would be an issue to the overall quality of the project if all my other main variables were null.

Another bad null result would be the coefficient on the effect of sales. While not my exact coefficient of interest, it would make it impossible to back out time fixed effects and cleanly identify the policy effect. It would not kill the project as I can do placebo tests and limit the time period after the policy change but it would harm identification.

Overall, it is important I start work on a few big games before expanding my sample to the universe of steam games. Then I can get an estimate of whether there are sizeable and significant effects to this policy. If not, then I do not waste too much time but otherwise I can begin the full scale project.

8 References

Schoenmueller, Verena and Netzer, Oded and Stahl, Florian, The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications (February 15, 2019). Columbia Business School Research Paper No. 18-10, Available at SSRN: https://ssrn.com/abstract=3100217 or http://dx.doi.org/10.2139/ssrn.3100217

Lin, D., Bezemer, CP., Zou, Y. et al. An empirical study of game reviews on the Steam platform. Empir Software Eng 24, 170–207 (2019).

Liyang Sun, Sarah Abraham, Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,Journal of Econometrics,Volume 225, Issue 2,2021,Pages175-199,ISSN 0304-4076,

Steam DB: https://steamdb.info/

Steam Store : https://store.steampowered.com/

Metacritic: https://www.metacritic.com/