The Impact of Online Reviews on Product Sales: Evidence From Video Games

Dmitry Sorokin (NYU) Ryan Stevens (NYU)

February 20, 2020

Abstract

Using high frequency data from a large video game marketplace we study the causal effect of reviews on sales. First, by restricting attention to products with unlikely quality changes and controlling for pricing behavior of firms, we make the case that panel data methods could be sufficient for identification. Second, we use a regression discontinuity approach, lately developed in the literature, to provide quasi-experimental evidence on the subject. We uncover an identification issue with this approach, and suggest a modified version, that significantly moderates the effects. We find that having better reviews increases sales by 3-7%, that the effect is stronger for games with more reviews and later in their life cycle.

1 Introduction

Online retail and search relies heavily on user-generated reviews. Increasingly folk wisdom promotes the positive impacts from higher reputation on firm outcomes. Anybody who has ever shopped online would probably agree that reviews matter at least somewhat for purchasing decisions. However, quantifying that impact of reviews on purchases is a difficult task. First, it is not obvious if simple review statistics, like the number of stars a product has,

matter for purchasing decisions (Dai et al. 2018). Customers could, instead, place a bigger weight on individual reviews that resonate with them. Second, even if the effect is there, in any setting there are plausible omitted variables that could lead to both higher ratings for a product and higher sales, obscuring the effect. For example, in restaurant markets, a menu change could lead to better food, which would cause both reviews and sales to trend upwards. In this study we quantify the causal effect of review statistics on purchases in a new market. Using panel data methods we find that increasing the fraction of positive reviews by 10 percentage points increases sales by 1-4%. Using a regression discontinuity approach we find that promoting a game to a better review category (for example, from "Mostly Positive Reviews" to "Positive Reviews") can increase sales by as much as 8%.

Our study provides three main contributions to the literature. First, we extend the analysis of review systmes to an important marketplace, an online computer video games platform *Steam*. Total revenue in the U.S. video games market has reached $43.4 billion in 2018, matching the size of the U.S. film industry (Minotti 2019), making it a worthwile research object. Steam is fundamentally similar to other platforms where users can both leave reviews and buy goods and that were previously studied in the literature: Amazon, Barnes and Noble, Trip Advisor, etc., which means that the results obtained for this market could be relevant in other settings. We employ an identification approach that has previously only been used in studying Yelp.com, a restaurant review aggregator, which is not a marketplace. Thus, to the best of our knowledge, we are the first paper to use a quasi-experimental approach in the setting of an online marketplace.

Another advantage of focusing on the video games market is the richness of the data. Our second contribution is leveraging these newly available data to address issues present in previous studies. As mentioned previously, the major challenge in eliciting the causal effect of reviews on outcomes is the presence of unobserved factors, such as quality, that affect both reviews and those outcomes. A related problem is firms' responses to review changes, e.g., marketing campaigns or price changes launched in response to review changes (see, e.g.,

Hollenbeck, Moorthy, and Proserpio 2019). To deal with the quality issues, we select a specific subset of games whose quality is fixed. Games, being digital products, are more likely to have a fixed quality, unlike, for example, restaurants, or manufactured goods. We validate our sample selection approach with production notes left by developers indicating when updates to the game were made, and show that products in our sample have very few updates over their life cycle. The high frequency of our reviews and player activity data minimizes the risk of pickig up the effects of unobserved firm actions, as we are able to estimate the (near) instantaneous impact of reviews on purchases. In addition, we observe the entire price history of each game, which allows us to control for a key lever that firms may use to respond to changes in their reviews. We show that discounts are a large positive driver of player activity, so controlling for promotions is important if they are correlated with review shocks.

Third, we highlight and propose a solution to a bias that arises in regression discontinuity (RD) designs where the running variable is a fraction (such as percentage of positive reviews among all reviews). This design has been used in the literature to study the effect of reviews on product outcomes (Anderson and Magruder 2012; Luca 2016). Consider the following stylized example to better understand our finding. Imagine a review system that allows users to leave a positive or a negative review for a product. If at least 50% of the reviews are positive, the product is labeled as "Good", and it is labeled as "Bad" otherwise. The existing approach in the literature is to take, say, products with 48-49.99% of positive reviews, and to compare them to products with 50-52% of positive reviews. The idea is that these products must be similar, on average, but the latter group (quasi-randomly) received a better label from the platform, so the difference in outcomes between the two groups can identify the average treatment effect of having "Good" reviews.

Imagine now that there are also two types of products: small and large. The smalll ones have, on average, 10 reviews and the large ones have a 100 reviews, on average. Among the large ones, products with 48 and 49 positive reviews would make it to the "Bad" part of the discontinuity, and products with 50-52 positive reviews would make it to the "Good" part.

Now, the problem with this design is that the small product can only make it to the "Good" part of the discontinuity: 5 positive reviews out of 10 is exactly 50%, but 4 and 6 positive reviews place the game outside of the RD bandwidth. As a result, games on different sides of the cutoff are, by no means, similar. The estimated effect of having a positive review label on, say, sales, in this example would be biased downward, as high sales of large games with "Good" reviews will be averaged with low sales of games with "Good" reviews, while the comparison group would only consist of large games with "Bad" reviews. Another problem with this design is that a large game with 49% of reviews being positive might need a substantial amount of new reviews to cross the 50% cutoff, which makes questionable the assertion that this game is similar to a large game with 51% of positive reviews. To deal with this bias, we propose a different discontinuity approach. Instead of asking how many percentage points a game needs to cross to a different review bin, we look at the total number of reviews the game needs to cross. We find that going from percentages to counts signficantly lowers the effect of reviews on sales. Thus, our paper suggests that results obtained in other papers employing the aforementioned regression discontinuity design could be biased upwards.

The rest of the paper is organized as follows. In Section 2 we present a brief literature review detailing the connection between this study and the literature. In Section 3 we introduce our institutional setting, as well as, our data and sample selection procedure. Section 4 details our empirical strategy and results, and Section 5 introduces our new regression discontinuity approach. We conclude in Section 6.

2 Literature

We contribute to the literature on using quasi-experimental variation to estimate the impact of reviews on purchases[1]. Early work using difference-in-difference estimates began with

[1]There exists a long literature of observational studied trying to estimate the impact of reviews on purchases. Given that our article is focused on a specific form of quasi-experimental variation to identify the impacts of reviews, we focus on these articles. For a review of the older literature consult Zhu and Zhang (2010), and the newer literature is covered by Luca (2016)

Chevalier and Mayzlin (2006) studying the impact of book reviews on purchases on bn.com and Amazon.com. Using relative differences in reviews across these two sites to control for product fixed effects, the authors find a positive relationship between reviews and sales ranks. Zhu and Zhang (2010) use similar methods to estimate the impact of reviews on purchases in the console video games market. While their focus is on moderating factors between reviews and purchases, the area of study (video games) and the study design (quasi-random experiments) are germane to our paper.

A problem with these difference-in-difference approaches is two fold. First, there is often little variation between reviews across sites (or consoles) leading to efficiency issues in estimating effect sizes. This can be seen in Zhu and Zhang, where the impacts of reviews on purchases is relatively small. Second, these approaches fail if there are good-platform specific effects (Zhu and Zhang 2010). If certain books (or games) have different promotional activity, population, or other demand side factors on a site (or a console), then this invalidates the difference-in-difference assumptions.

To handle these concerns, recent articles use regression discontinuity approaches. Specifically, Luca (2016) and Anderson and Magruder (2012) study the impact of half-star changes to restaurants' reviews on the platform site Yelp. Yelp computes average ratings on a scale from 1 to 5 and awards stars from 1 to 5 in half-star increments. These star awards form a natural discontinuity. For example, a restaurant with an average rating of 3.26 will get 3.5 stars, whereas a restaurant with an average rating of 3.24 will get 3.0 stars. Luca combines review data with revenue data from the city of Seattle to estimate the impact of Yelp half-stars on revenues. Using a within estimator, controlling for restaurant and time fixed effects, Luca finds 5-9% revenue gains from a half-star change in review score, mainly driven by non-chain affiliated restaurants (Luca 2016). Anderson and Magruder use the same discontinuity, focusing on a different city, San Francisco, and a different outcome variable, the ability to book a table 36 hours in advance. Using a pooled panel estimator, controlling for distance to the threshold, they find a negative relationship between stars and ability to

book a table (Anderson and Magruder 2012).

Our paper also provides a method to determine the proportion of buyers that leaves reviews. Given the public goods nature of reviews, there is a large liteature studying motivations influencing consumers decisions to write reviews (Burtch et al. 2018). However, there is surprisingly little research attempting to estimate how many purchasers actually leave reviews. Prior work that has provided an estimates of review prevalence have access to internal company data. For example, Fradkin et al. observes approximately 120K transactions over a single month, with 67% of guests and 72% of hosts leaving reviews (Fradkin, Grewal, and Holtz 2019). Another study focuses on early eBay in 2010 and shows that in ~47% of ~400K transactions observed for a sample period in 1999, buyers left no feedback (Jian, MacKie-Mason, and Resnic 2010). While these papers provide exact numbers for the conversion rate from purchase to review, they are limited to two-sided review systems, where both a buyer and a seller can leave a review. Additionally, most researchers do not have access to these internal numbers. In the appendix, we provide our simple method to impute conversion rates using a proxy for purchases.

3 Data

3.1 Institutional Setting

The object of our study is the review system of *Steam*—the largest digital marketplace for selling PC video games. Exact data on Steam's market share is hard to come by, but in 2013 it was responsible for about 75% of PC games sold online (Cliff 2013). From 2009 to 2018, there was a sharp decline in the importance of traditional physical distribution chanels for video games: the share of games sold offline went from 80% in 2009 to 17% in 2018 (Association 2019). The video games market itself has seen tremendous growth in recent years. Total revenue in the U.S., which includes purchases on PC's, consoles, and smartphones, has reached $43.4 billion in 2018, matching the size of the U.S. film industry

(Minotti 2019). Thus, Steam is a major player in a large and growing market for video games. Online reviews is a defining feature of online commerce, thus understanding Steam's review system is important for understanding the functioning of this vibrant market.

Steam introduced its review system in November 2013, replacing the existing Steam Recommendations service ("Steam Reviews Now in Beta" 2013). The system was (and still is) constantly developing, with a large modification of the system happening at the end of 2016. For that reason, and due to data availability issues, we will focus on years 2017-2019. During the years under study, the review system functioned as follows. Any user who purchased a game on Steam could leave a review for the game, and that review would contribute to the game's visibility. A review writer has to assign a binary grade ("Lke" or "Dislike") to the game, and to write at least a short review of the game. Once a game has 10 reviews, it is assigned a *review score* and a *review bin*. The review score is defined as a percentage share of positive reviews ("likes") among all reviews, rounded down to the closest integer. For example, a game that has 14 positive and 2 negative reviews would have a review score of 87%.

Review bins are verbal assessments of the quality of game's reviews, that are assigned by Steam based on the review score of the game and the number of reviews a game has. The lowest bin characterizes the reviews of a game as "Overwhelmingly Negative", and the highest one as "Overwhelmingly Positive". In our analysis we will focus on 3 intermediate consecuitive review bins: "Mixed", "Mostly Positive", and "Positive". A game is labeled as having "Mixed" reviews if its review score is between 40% and 69%. It is convenient and fairly innocuous to characterize all games with a lower score as having "Negative" reviews. Once the game crosses the 70% score cutoff, it becomes labeled as "Mostly Positive". It is the first "good" review bin: the tag is displayed in a pleasent blue color, as opposed to the orange color of the "Mixed" tag and the red color of all the negative bins. Once the game crosses the 80% cutoff, it is labeled as either "Positive" or "Very Positive", depending on the number of reviews the game has. We will blend the two together, and say that all games

with the review score exceeding 80% have “Positive” reviews.

The review score affects the placement of the game on Steam’s web pages. A user visiting Steam’s frontpage is welcomed with several small selections of featured games, and can choose among many ways to further browse Steam: she can browse by genre, look at games on sale, trending games, etc. If she chooses to browse, games satisfying her query will be presented to her in a list. Games that are rated “Mixed” or above receive a significant boost compared to games rated below on such lists (“Positive "Review Bombs"” 2019). The overall contribution of the review score to visibility for games is small relative to other factors, and doesn’t vary much across “Mixed”-“Positve” bins. However, hovering over the game shows the review bin of the game and the total number of reviews (see Figure 1), and, in some query responses, the raw review score as well. Therefore, conditional on the placement assigned to games by Steam, a better reviewed game could still attract more clicks, and, eventually, buyers. If a user decides to visit the page of a particular game, she can get more information about the game and about the reviews the game has. She will see the raw review score, most helpful and most recent reviews, as well as, get access to various filtering tools, such as the language of the review, playtime of the reviewer, etc.

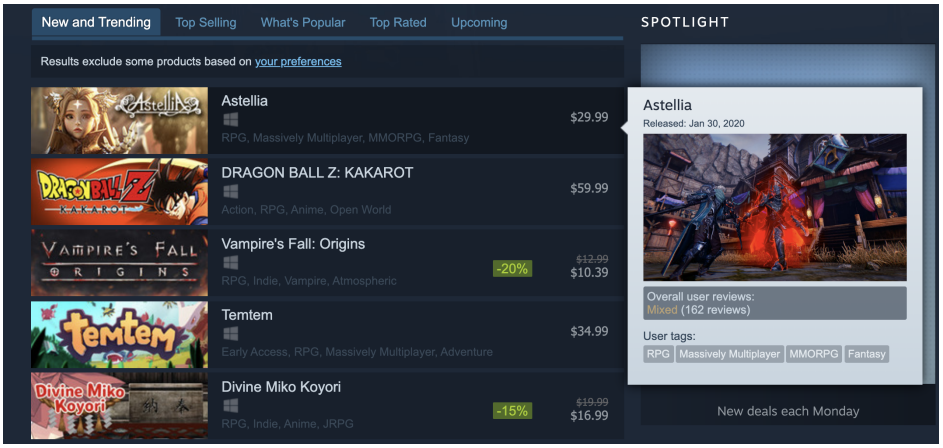

Figure 1: Example browsing session on Steam

3.2 Data Sources

The data we use comes from two sources, Steam and steamDB. On Steam, the reviews are publicly available. We gather individual reviews, and use the time stamps on the reviews to reproduce the exact state of reviews for any game on any day. Steam is committed to having a transparent review system, so the review score has always been calculated as a fraction of positive reviews among all reviews. The actual amount of information available to shoppers is more complicated than just review scores and review bins. Users can read the raw reviews, and the reviews are ranked based on their users-assigned helpfulness, but we can not reproduce all the details retroactively [2]. Therefore, in our analysis we only use the review score and the review bins as summaries of the review status of a game.

All other data comes from *steamdb.info*—a third-party website dedicated to providing real time information about games on Steam. It contains the universe of games published on Steam, including all the information that is available directly from Steam (release date, publisher, genre, and other "static" information about the title), as well as, data that can't be accessed on Steam: price history and concurent players history. The owners of SteamDB declined to disclose how they collect the data, but this website is extensively used by magazines writing on PC gaming, and presents the best available source of data on Steam games[3].

The important aspect of the data is its dynamic, panel nature [4]. We are able to not only observe the entire price history, but a proxy for purchases in real time. Specifically, we refer to the maximum concurent players variable as the *player count*. For every game, this variable measures the maximum number of people playing the game *simultaneously* on a given day[5].

For every game we also observe daily price data. As we will show later, prices are

[2]This is a general problem using scraped data to analyze review systems. For example, Yelp does not display reviews in the order they were written.

[3]Price data, in principle, could be collected by regular scraping the Steam store.

[4]All the static elements of a game, such as it's genre are differenced out in our 'within' regressions in the Analysis section. We could report analysis by genres, developer types etc, however, this is not the main focus of our study

[5]This variable underestimates the total number of players who play the game on any day because gamers within a day don't play the game at exactly the same time.

relatively sticky on Steam. Most games never change their price. The only variation in prices comes from discounts that firms can run according to the policies outlined by Steam. Our price variable measures the lowest price available on a given day. The shortest allowed duration of a discount on Steam is one day (Documentation, n.d.), so even the shortest discounts are observed.

3.3 Sample Selection

To select our sample, we have two major concerns: stability of the review system and changing quality of the games. In this section, we explain our sample selection procedure. First, we exclude all games that were released before January 1, 2017. Thus, our data spans two years, from January 2017 to January 2019. The major reason for this restriction is that the review system on Steam was massively overhauled in September of 2016. Growing concerns of gaming the system led to Steam's exlusion of a big fraction of reviews on the platform and changing the rules for leaving reviews. The period we are using is characterized by a more stable review system, and by restricting attention only to games that were launched after that change, we ensure that the data is not contaminated by big structural shocks to the system.

The major challenge in establishing a causal effect of online reviews on outcomes is that products with high unobserved quality can have both good outcomes and high reviews. Therefore, it is desirable to minimize the possibility of quality fluctuations for the products under study. To that end, we exclude all games that have an online multiplayer format. The quality of multiplayer games inherently depends on the network effects created by other users, and thefore introduces undesirable variation in quality. Multiplayer games exhibit a completely different usage pattern, that doesn't necessarily fade with age, because the developers keep improving the game if the interest persists, which further exacerbates the problem of changing game quality. We restrict attention to single player games that are more similar in nature to standard durable goods. The developer develops the product,

tests it, and releases a complete game that is ready for consumption. Each game like that has a bounded playtime, and potentially can be "completed", which makes it similar in nature to books or movies [6]. We also exclude games that have been through so-called "Early Access"—a program that allows developers to release partially completed games, with the goal of soliciting feedback from the community. By design, such games are bound to change their quality and could contaminate the results of the present study.

We validate our selection approach by checking that games in our sample publish fewer update announcements. On Steam,developers can post notes directed towards consumers highlighting updates to the game, as well as, other promotional material. These developer notes (or patch notes) are meant to convey potential changes in gameplay to consumers . We scrape these developer notes for all games in the period under study. Some of these notes simply highlight promotions, as well as, scheduled release of different content by the same developer. Because we are interested in changes to quality, we focus on notes that highlight bugs or isuses with the game. To do so, we classify a developer's post as a bug fix if it includes the stem words 'bug' or 'fix' in the text of the post. For each game released in 2017 or later, we count the number of these bug fix posts, as well as, all posts created by the developer. Figure 2 shows the median number of bug fixes and developer posts across games included in the sample and games excluded from the sample. Games that are not included in the sample (multiplayer or "Early Access" games) are more likely to have bug fixes than games included in the sample. In fact, the median number of bug fixes over the lifetime of the game in our sample is only 1.

The last big restriction we employ concerns game size. Steam, like many other platforms, exhibits very fat tales in the popularity of games (Prause and Weigan 2019). Out of approximately 22,000 games that we have data for, more than 18,000 had less than 5 people playing them on a median day. Prior work has shown reviews help firms with smaller review bases (Anderson and Magruder 2012). Becase of this, we don't want to focus exclusively on

[6]Some support and maintaining is still done to keep the games playable on newer operation systems, and we can not fully rule out the possibility that some quality improvements are made.

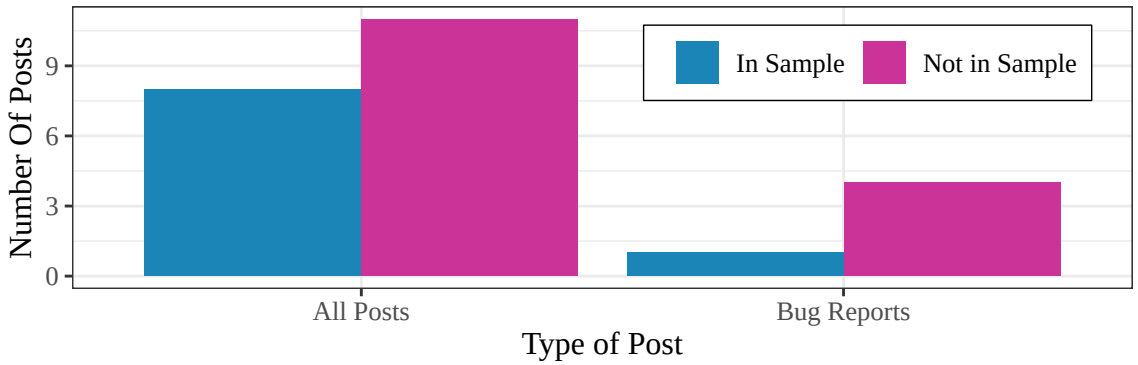

Figure 2: Number of developer posts and bug fixes in and out of sample

large games, but we deem the games in the tail to be too small. The quality of the player count data for such games is questionable. Therefore, our sample only includes games that had at least 5 people playing them on the median day.

3.4 Sample Description

The final dataset we employ is an unbalanced daily panel that consists of 893 games (entities, i) and spans 765 days (periods, t). All games in the sample are observed from their release date, so some games are observed for longer than the others. The distribution of time in the sample is

Table 1: Summary Of Time in The Sample

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Time in Sample	409.20	221.99	7	205	616	762

For every game i on any day t we observe the player count, the total number of positive and negative reviews the game had on that day, and the price. Figure 3 shows what this data looks like for an example "median" game in the sample.

As was mentioned before, the majoriry of games on Steam have a small player base, which also holds in our data. Figure 4 shows that half of the games had less than 17 concurrent players playing when the game was thirty days old.

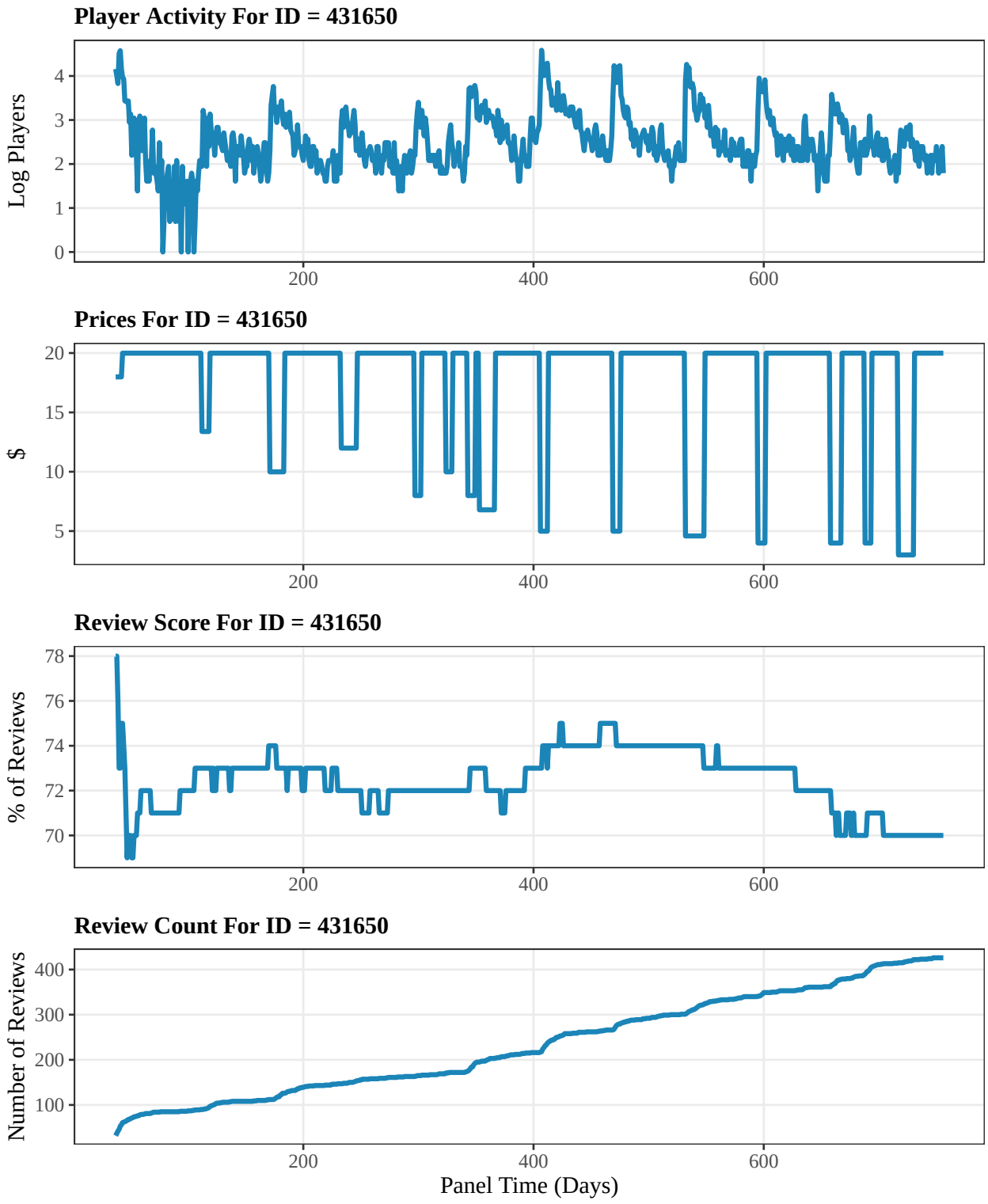

Figure 3: Example observation of a game in the sample

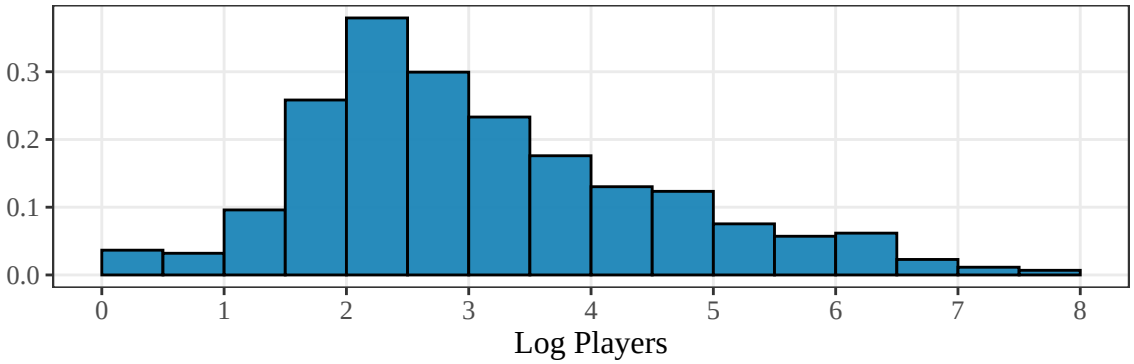

Figure 4: Players at the age of 30 days

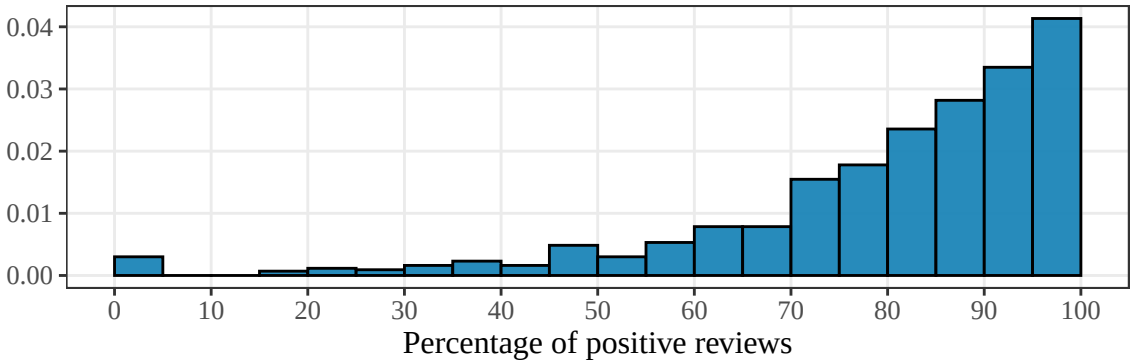

Figure 5: Distribution of review scores at the age of 30 days

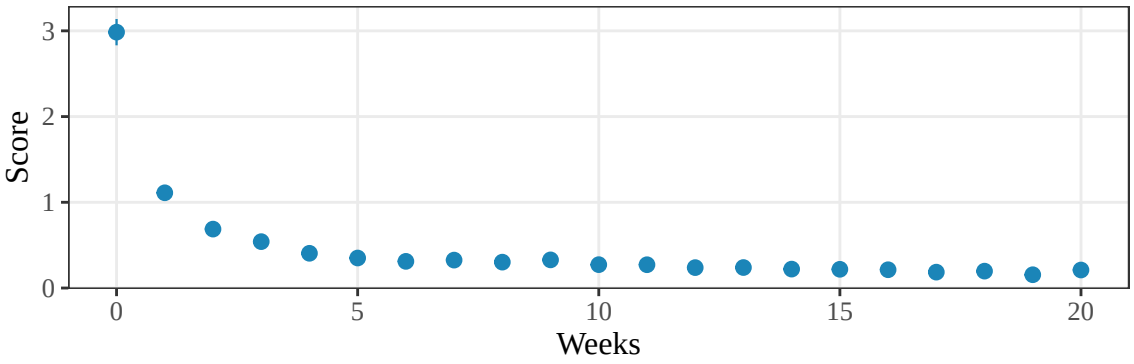

Figure 6: Within Week Standard Deviation In Review Score, Averaged By Games

Games don't change posted prices very often. We only observe 160 incidents of price change. Instead, games go on discounts. We observe 6296 individual discounts, which can last anywhere between one day and two weeks. The price path in Figure 3 represents well the standard price behavior in the sample.

Central to our analysis is heterogeneity in reviews. Being able to identify when a product was released is important to determine not only *if* reviews matter for purchases, but *when* in a product's lifecycle do reviews matters. Figure 5 shows that the review distribution is extremely skewed. The vast majority of games have a review score above 60%. It is important to investigate how sticky the review score is, because lack of variation in the score could cause problems for identification. When the game just starts out, the total number of reviews is small, and a marginal review can change the review score a lot. As the number of reviews increases, it becomes harder to affect the score. Figure 6 shows that the variance of the review score decays rapidly over time. However, there is still some variation even on the weekly level, and we are able to find robust effects for older games in our analysis.

An important source of heterogeneity in online markets is the number of reviews a product receives. Given that we know when a game is released, we observe the total stock of reviews over time. Games receive a lion's share of their total reviews in the first week. However, consumers continue reviewing the game even when the total number of reviews is already quite high.

Table 2: Total Number of Reviews By Age

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
First Week	127.07	227.14	0.00	13.00	135.00	1,920.00
First Month	190.04	335.44	0.00	19.00	200.00	2,733.00
Half a Year	314.24	530.46	0.00	38.00	328.00	3,757.00
First Year	438.39	718.17	1.00	53.00	438.00	4,170.00

4 Analysis

The main question we ask in this paper is whether reviews have a causal effect on purchases. The fundamental identification problem is that products with better quality are more likely to have better reviews. We address this identification issue in three different ways.

First, our sample has been selected to minimize the presence of games whose quality might change over time. If quality is unobserved but stable over time, standard panel data methods are enough to identify the causal effect of reviews on outcomes. However, if there remain important time-varying unobserved factors that influence reviews and player activity (or purchases), then panel methods would fail at the task.

Second, we follow the literature and exploit a discontinuity present in many markets that employ review systems. Each game on Steam is assigned a label that summarizes the reviews the game has recieved: "mixed reviews" if the fraction of positive reviews exceeds 40%, "positive reviews" if the review score exceeds 70%, etc.. We compare games on both sides of such thresholds. Identification is obtained if games sort quasi-randomly around them, i.e., if one game just happened to get enough reviews to cross the cutoff, and the other one did not. We identify a novel challenge with this approach that is relevant for all studies relying on such discontinuities, and propose a proceedure that solves the problem.

We believe that the two approaches we use complement each other. We find that a 10 pp change in the review score increases purchases by 1-4%, on average, and that upgrading to a better review bin icnreases purchases by 2-8%. Robustness exercises support an effect of 2-4% for some review bins.

4.1 Evidence From Panel Data

Our empirical strategy is derived from the following model of the observed player count:

$$g_{it} = \psi g_{it-1} + \lambda_{it} + \mu_i + \tau_t + \epsilon_{it} \tag{1}$$

Here g_{it} is the log player count for game i on day t, λ_{it} is the inflow of new players, μ_i is an unobserved game specific effect (e.g. quality of the game), and τ_t is a set of time effects: day of the week effects and calendar week effects. The idea behind the model is that people playing the game consist of continuing players and new players. New players are the recent buyers of the game, while continuing players have purchased the game in the past and keep playing until they are bored with the game or have completed it. Thus, $1 - \psi$ measures the rate at which existing players stop playing the game.

The new players of the game, λ_{it}, are the recent purchasers of the game. We assume that λ_{it} is a linear function of the variables that matter for purchasing decisions: discounts, review score or bins, and age of the game. Discounts are big drivers of player counts in the data, so it is an important control variable. Age is included to allow the demand for the game to follow some time trend. The coefficients on the review variables measure the effect of reviews on demand, and are the main coefficients of interest.

Games differ in their unobserved quality μ_i, which could simultaneously lead to a higher inflow of new players (purchases) and to better reviews. Time effects τ_t include dummies for days of the week and dummies for calendar weeks. The former are important to capture the within-week seasonality that is present in the data: people play (or buy) more often during the weekends. The latter allow us to control for platform-wide shocks like holidays, when people play more.

The final specification we estimate is

$$g_{it} = \psi g_{it-1} + \alpha \log(price_{it}) + \beta scoreBin_{it} + \gamma age_{it} + \mu_i + \tau_t + \epsilon_{it} \tag{2}$$

The inclusion of the autoregressive component is dictated by the nature of our data. We don't directly observe the variable of interest (purchases), and therefore we need to make this extra step to be able to identify the effect of reviews on purchases. Some regressors that are potentially important, like the genre of the game or the identity of the developer or publisher

didn't make it to the regression, because those variables remain constant throughout the entire life-cycle of the game, and thus would be differenced out in the fixed-effect estimation.

It is known since Nickell (1981) and Anderson and Hsiao (1982) that the inclusion of the lag of the dependent variable as a regressor makes the within estimator of the parameters of the model inconsistent. However, the strength of the bias is inversly-related to the length of the time dimension. A game in our daily two year long panel is observed, on average, for 409 periods, so we proceed with the within estimator. A related problem is that the correlation that is introduced to the error term by the within transformation biases the inference. The size of this bias also decays with T (Anderson and Hsiao 1982). Therefore, everywhere in the paper we report White's heteroskedasticity-robust standard errors that allow for an aribtrary variance of the residual within the game and assume no serial correlation between ϵ_{it}'s (see, e.g., Akhmedov and Zhuravskaya 2004 for an example of the same approach with about 80 time periods).

4.1.1 Results

The results of estimating (2) are presented in Table (3). The first column is estimated on the entire sample. The second column restricts attention to the first year of each game's life, and the third one—on the first month. The reference group for the analysis are games with a "Mostly Positive" review label.

The results indicate that reviews have a strong effect on daily purchases. Relative to being a "Mostly Positive" game, having the "Negative" review label decreases purchases by 7.01%. This effect is even larger for games within their first year. The regression picks up an even stronger penalty in the first month. While prior studies have found a positive bump on purchases, we are able to show that this effect is larger for newer versus older products, which has impacts for platform design. Similarly, games in the "Mixed" review bin have purchases that are 4.65% lower. Interestingly, the effect is smaller in the first year, and even smaller in the first month (and not significant). Having "Positive" reviews, on the other

Table 3: Effect of review bins on sales—panel regressions

	Dependent variable:		
	Log Players		
	Full Sample	1st Year	1st Month
Lag Players	0.808^{***}	0.793^{***}	0.620^{***}
	(0.002)	(0.002)	(0.014)
Log Price	-0.108^{***}	-0.135^{***}	-0.300^{***}
	(0.004)	(0.006)	(0.045)
Negative	-0.070^{***}	-0.089^{***}	-0.148^{*}
	(0.016)	(0.021)	(0.085)
Mixed	-0.047^{***}	-0.028^{***}	0.007
	(0.007)	(0.008)	(0.029)
Positive	0.031^{***}	0.029^{***}	0.020
	(0.005)	(0.006)	(0.019)
Age	0.0004^{**}	0.0005^{**}	-0.015^{***}
	(0.0002)	(0.0002)	(0.001)
Observations	362,181	255,185	26,743
R^2	0.720	0.694	0.625
Note:			*p<0.1; **p<0.05; ***p<0.01

hands, provides a boost of 3.06%, and the effect also becomes stronger over time. The high estimated value of the autoregressive coefficient ψ supports our approach of controlling for continuing players. Significance of the price coefficient in all regressions we run in this paper documents the big role that discounts play in this market. Our estimate suggests an average elasticity of demand of 0.11 to 0.3[7].

Whether consumers should trust the review label they see could depend on the number of reviews the game has. We reestimate model (2) by the terciles of the number of reviews that games had accumulated by the age of one year. The sample only includes one year of data for all games that reached that age. Games at the bottom of the review count distribution, i.e. the ones having less than 79 reviews at the age of one year, don't exhibit

[7]Estimation of the elasticity of demand for video games is not the focus of this paper, but the estimates we obtain could be regarded as such even without relying on any instruments. The majority of games never change prices outside of discounts, so even if discounts are given when demand is low (or high), the immediate spike in the player counts identifies the slope of the shifted demand curve. Of course, video games are durable goods, so the implied elasticity would measure a particular response of some subset of, potentially, forward looking consumers to the price path that follows Steam rules.

Table 4: Effect of review bins by number of reviews

	Dependent variable:		
	Log Players		
	T1	T2	T3
Lag Players	0.748^{***}	0.785^{***}	0.852^{***}
	(0.005)	(0.004)	(0.004)
Log Price	-0.033^{***}	-0.194^{***}	-0.326^{***}
	(0.009)	(0.009)	(0.010)
Negative	-0.108	0.122^{***}	-0.036
	(0.068)	(0.024)	(0.027)
Mixed	-0.023	-0.039^{**}	-0.043^{***}
	(0.018)	(0.018)	(0.014)
Positive	0.021^{*}	0.038^{***}	0.040^{***}
	(0.011)	(0.010)	(0.014)
Age	0.001^{***}	-0.001^{***}	-0.0001
	(0.0002)	(0.0001)	(0.0002)
Observations	61,617	62,041	61,669
R^2	0.622	0.712	0.849
Note:			*p<0.1; **p<0.05; ***p<0.01

a strong difference between different review bins, except for the "Positive" one. However, the magnitudes of the point estimates are reasonable. Games at the top of the distribution, i.e. ones with more than 300 reviews, seem to be driving the effects we find in Table 3. The coefficient on the "Negative" bin in the second column is a puzzling outlier, but the rest of the column has effects of a similar size. Results in Table 4 suggest that the effect of reviews is stronger for games that have more reviews, which is in line with the idea that customers value more information. These results could also explain why the effects we found in Table 3 become stronger over time: as the game ages, it accumulates more reviews. Together these findings suggest that, even though reviews should be important in the early days of a game's life, consumers are cautious to attribute too much weight to review scores that are based on few reviews.

As a robustness, we also estimate the regression model (2) using the continuous review score variable instead of the binary review bin indicators. The results are presented in Table

(5). The first column is estimated on the entire sample, the second column restricts attention to the first year of each game's life, and the third one—on the first month. The effect of the review score on the number of players is still positive and statistically significant. A 10pp increase in the score, on average, increases the player count by 1.25%, with a higher effect of 3.69% in the first month. These effects are more moderate compared to the results in Tables 3 and 4. One explanation could be that a marginal point added to the review score just doesn't matter much, because the consumers don't distinguish between games with, say, a 74% and a 75% review score. A notable difference between the results is that Table 5 suggests that the effect of reviews decays with age, while Table 3 suggests the opposite.

Table 5: Effect of review score on sales–panel regressions

	Dependent variable:		
	Log Players		
	Full Sample	1st Year	1st Month
Lag Players	0.809^{***}	0.794^{***}	0.621^{***}
	(0.002)	(0.002)	(0.014)
Log Price	-0.108^{***}	-0.135^{***}	-0.302^{***}
	(0.004)	(0.006)	(0.045)
Score	0.001^{***}	0.002^{***}	0.004^{***}
	(0.0003)	(0.0004)	(0.001)
Age	0.0004^{**}	0.0005^{**}	-0.015^{***}
	(0.0002)	(0.0002)	(0.001)
Observations	362,181	255,185	26,743
R^2	0.720	0.694	0.626
Note:			*p<0.1; **p<0.05; ***p<0.01

4.2 Evidence From Regression Discontinuity

Another approach to eliciting the causal effect of reviews on product outcomes is to exploit the rounding of the review score employed by the platform (Anderson and Magruder 2012; Luca 2016). This approach is fully implementable in our context as well. Recall that a game has "Mixed" reviews when the score is between 40% and 69%. The score between 70% and

79% is labeled as "Mostly Positive". A game gets a lable of "Positive" or better when the score exceeds 80%. Anderson and Magruder (2012), in a setting similar to ours, estimate the following model for a cutoff R (say, 80%):

$$y_{it} = \beta_0 + \beta \mathbb{1}(score_{it} \geq R) + \beta_1(score_{it} - R) + \beta_2(score_{it} - R)\mathbb{1}(score_{it} \geq R) + \epsilon_{it},$$

focusing on observations close to the cutoff. We replicate their results here, using $y_{it} = g_{it}$. Imbens and Lemieux (2008) suggests that control variables can improve the accuracy of the regression discontinuity regression, so we employ the same set of controls as we used in our panel regressions: lagged players, price, age, game effects, day of the week effects, and panel week effects. We will refer to $\mathbb{1}(score_{it} \geq R) = T_{it}$ as "treatment", and the respective coefficient is the coefficient of interest. Coefficients on the control variables don't have a causal interpretation. Our final specification is

$$g_{it} = \beta T_{it} + \beta_1(score_{it} - R) + \beta_2(score_{it} - R)T_{it} + \psi g_{it-1} + \alpha \log(price_{it}) + \gamma age_{it} + \mu_i + \tau_t + \epsilon_{it} \quad (3)$$

4.2.1 Results

Table 6 presents the results of estimating equation (3) (we omit the coefficients on the control variables for brevity). We use a bandwidth of 3 percentage points around three cutoffs: 40% ("Mixed"), 70% ("Mostly Positive"), and 80% ("Positive"). The results suggest that crossing the "Mixed" and "Mostly Positive" thresholds leads to an 7-8% increase in purchases. The effect of crossing the "Positive" threshold is smaller. The results for "Mostly Positive" and "Positive" bins are robust to smaller bandwidths of 2 and 1 p.p. around the cutoffs (see the robustness section).

One way to assess robustness of the regression discontinuity analysis is by inspecting the picture of the raw data. We apply the within transformation to the data to difference out game and time effects, and plot the normalized average player counts by distance (in

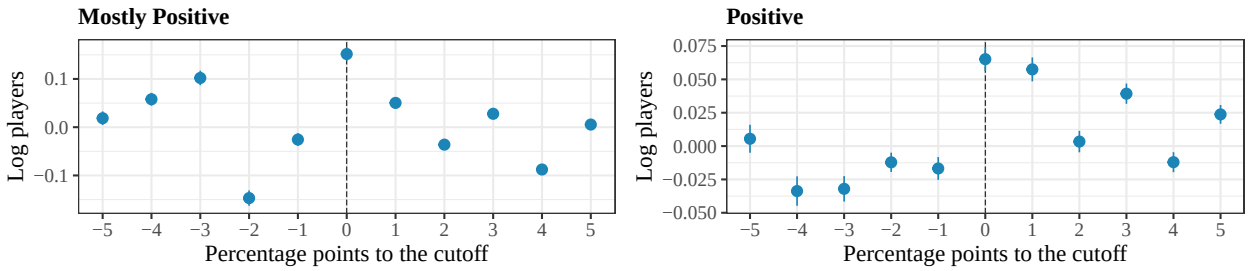

Figure 7: Discontinuity in player count at different cutoffs

percentage points) to the cutoff. The results are displayed in Figure 7. The picture for the "Positive" threshold exhibits a more pronounced discontinuity, while the effect for the "Mostly Positive" bin is mostly driven by the games exactly at the cutoff.

Table 6: Regression discontinuity results

	Dependent variable:		
	Log Players		
	Mixed	Mostly Positive	Positive
Score-R	-0.039^{**}	-0.016^{**}	0.010^{**}
	(0.015)	(0.007)	(0.004)
Treatment	0.077^{**}	0.072^{***}	0.024^{***}
	(0.035)	(0.015)	(0.009)
(Score-R)*T	0.004	0.008	-0.017^{***}
	(0.022)	(0.008)	(0.005)
Observations	9,384	32,502	54,479
R^2	0.636	0.652	0.678
Note:			*p<0.1; **p<0.05; ***p<0.01

5 Validity Of The RD Approach

The validity of the regression discontinuity approach in our setting varrants a discussion. What distinguishes our setting from classic RD settings is the highly dynamic nature of the data. Potentially, games in our sample can be crossing the review bin thresholds multiple times in short periods of time. Suppose that, once the game is upgraded to a higher bin, it

gets a permanent boost to the player count, that it never loses even if it crosses back to a lower bin. A reason could be that, a game becomes exposed to a bigger audience, who then continue to recommend the game through offline chanels. If that is the case, games around the threshold would have similar player counts, because all these games have been above the threshold at least once, and the effects of crossing would be washed away. Such concerns, however, could only explain the lack of the effect, while we do find an effect. In the robustness section we repeat the exercise excluding games that cross the threshold multiple times within a certain time frame, and, in line with the logic outlined above, our results become stronger.

Anderson and Magruder (2012) brings up another point, noting that the firm can respond to crossing the threshold by changing their marketing or pricing behavior, thus contaminating the causal effect of having a higher rating with the effect of, say, marketing polcies. We can't rule out a marketing intervention, but, unlike that paper or Luca (2016), we have access to pricing data, so we control for these effects in our regressions.

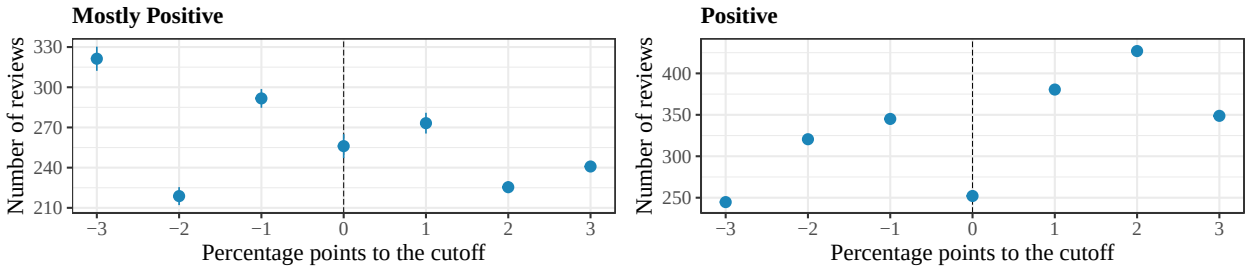

Figure 8: Discontinuity in review count at different cutoffs

We identify a novel issue with using the regression discontinuity approach to study causal effects of reviews. A fundamental assumption of RD is that, upon crossing the threshold, treatment variable is the only variable that changes discontinuously, and other characteristics change smoothly. We claim that this assumption could be violated mechanically in settings similar to ours. Consider a threshold of 50%. To have a review score of 50% a game could have 50 positive reviews out of 100, or, alternatively, it could have 25 positive reviews out of 50. However, between the two games, only the first one could instead find itself 1 percentage

point above or below the cutoff, if it gets an extra review. The game that only has 50 reviews to start with, can never be found 1 percentage point away from the threshold, as 26 positive reviews would increase the score of this game to 52%, entirely skipping 51%. Thus, the effect of crossing the threshold could, in part, be contaminated by the difference in review numbers across games that are at the threshold and games some number of percentage points away from them. This type of of issue is not restricted to only our context. For example, in studies using outcomes from elections, those elections with more voters will be closer to the threshold than those elections with fewer voters. Election contests with more voters may coincide with denser and more urban locations than those places with fewer voters (Lee 2008).

Figure 8 demonstrates that, in our sample, this is the case for the "Positive" threshold[8], but not for the "Mostly Positive" one. Close examination of Figures 7 and 8 shows that our results for the "Positive" cutoff are not fully driven by games exatcly at the cutoff, but this observation, nevertheless, raises concerns about the validity of the RD approach to study the impact of reviews on product outcomes.

5.1 Discontinuity in Review Count

In light of the identification problem outlined above, we propose a novel discontinuity approach that doesn't suffer from the mechanical discontinuity of other covariates around the review cutoffs. Instead of asking how many percentage points a game needs to upgrade to a better review bin, we ask how many *reviews* it needs to upgrade. This approach is much more in the spirit of the classic regression discontinuity design philosophy: we propose to compare two similar games, one of which just happened to get a couple of extra reviews that pushed it to a better review bin, while the other did not. We estimate the same specification,

$$g_{it} = \beta T_{it} + \beta_1(score_{it} - R) + \beta_2(score_{it} - R)T_{it} + \psi g_{it-1} + \alpha \log(price_{it}) + \mu_i + \tau_t + \epsilon_{it}, \quad (4)$$

[8]Not shown here, but it is also the case for the "Mixed" threshold

except that, instead of focusing on games with a score that is within a 3 p.p. of the cutoff, we use games that have extra (or lack 3 extra) reviews to clear the bar.

Table 7: RDD in the number of reviews

	Dependent variable:		
	Log Players		
	Mixed	Mostly Positive	Positive
Score-R	−0.001	0.002***	−0.005
	(0.001)	(0.001)	(0.003)
Treatment	−0.014	0.041***	0.027*
	(0.020)	(0.014)	(0.014)
(Score-R)*T	0.003	0.002	0.0002
	(0.002)	(0.003)	(0.007)
Observations	26,836	25,094	16,557
R^2	0.703	0.674	0.581
Note:		*p<0.1; **p<0.05; ***p<0.01	

The results (omitting the controls) are presented in Table 7. Compared to the RD results from Table 6, we see no effect of crossing the "Mixed" threshold at all. The effect of crossing the "Mostly Positive" cutoff is almost twice smaller, and the effect for the "Positive" one is similar, albeit less significant. Remarkably, the effects here are very close to the effects obtained from the panel analysis in Table 3.

It is instructive to compare the number of observations used for different bins across these two tables: Table 6 has more observations for higher review bins, while the opposite is true for 7. The reason is that, in the data, a game is more likely to be 3 reviews away from getting "Mixed" reviews than it is from turning "Positive", because better reviewed games tend to have more reviews. A game with a review score of 77% would be used for our discontinuity in *score* regression, but would most likely not make it to the discontinuity in the *number* of reviews regression, because a median game like that is 16 reviews short of turning "Positive", with the largest gap we observe in the data being 211 reviews. To put it another way, the amount of additional positive reviews a median game with a 77% review score needs to cross the "Positive" threshold constitutes around 14% of its total review count.

This observation questions the validity of the discontinuity in scores approach, because games on different sides of the cutoff are, in a sense, very far away from each other. Our results in Table 6 are robust to using really small bandwidths of 1 and 2 p.p., but it is not uncommon in the literature to rely on even larger bins (Anderson and Magruder 2012, e.g., in their study of reviews on Yelp uses what would be a 5 p.p. bindwidth in our setting.)

6 Robustness

6.1 Multiple Crossings

An important critique of using a regression design framework, discussed earlier, is the fact that products can migrate between several review bins over time, and, therefore, wash out the effects we are trying to uncover. We define "multiple crossings" as when a game crosses a threshold twice within a 14 day window. We reestimate (3), both for RD in scores and RD in number of reviews, dropping games that have "multiple crossings". The results are presented in Table 8. Compared to our baseline results in Tables 6 and 7, these effects are uniformly larger. This finding confirms that the baseline results are conservative and that multiple crossings could undermine the validity of regression discontinuity regressions in dynamic settings.

Table 8: RD results excluding games with multiple crossings

	Mixed		Mostly Positive		Positive	
	S	N	S	N	S	N
Treatment	0.082**	−0.007	0.086***	0.047***	0.024***	0.026*
	(0.036)	(0.020)	(0.016)	(0.015)	(0.009)	(0.014)
Observations	9,181	26,290	31,730	24,107	54,426	16,399

Note: *p<0.1; **p<0.05; ***p<0.01
RD in: S = scores, N = numbers

6.2 Binwidth Robustness

Table 9 presents the results of estimating model 3, our RD in scores specification, for smaller neighborhoods around the cutoffs: 2 and 1 percentage points. Results for "Mostly Positive" and "Positive" are extremely robust. We get a negative coefficient for "Mixed" at the binwidth of 2 p.p., and the coefficient at 1 p.p. is sizably larger than the one found in Table 6, so the results for "Mixed" are less robust.

Table 10 presents the results of estimating our RD in review count specification for different cutoffs as well. In the baseline we used a distance of 3 reviews from the threshold. In the robustness exercise we use 4 and 2. The coefficient on "Mostly Positive" is the most stable one, followed by "Positive". The results for "Mixed" were not significant in the baseline, and remain not significant or negative here.

Table 9: Robustness check for RD in review scores

	Mixed		Mostly Positive		Positive	
	+/-2	+/-1	+/-2	+/-1	+/-2	+/-1
Treatment	−0.114**	0.124**	0.067***	0.064***	0.022*	0.028**
	(0.050)	(0.061)	(0.020)	(0.021)	(0.012)	(0.013)
Observations	7,012	4,531	23,475	13,155	38,286	22,414
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 10: Robustness check for RD in review counts

	Mixed		Mostly Positive		Positive	
	+/-4	+/-2	+/-4	+/-2	+/-4	+/-2
Treatment	0.006	−0.054***	0.038***	0.001	0.026**	0.019
	(0.019)	(0.021)	(0.013)	(0.018)	(0.012)	(0.026)
Observations	32,736	21,268	30,469	16,829	22,159	10,290
Note:						*p<0.1; **p<0.05; ***p<0.01

7 Conclusion

This paper studies the causal effect of review statistics on sales on a major online marketplace for computer games. We take advantage of the high frequency of our data, the availability of detailed price and usage data, the digital nature of the product of study, and the institutional details to execute several alternative approaches to identifying the effect of interest.

First, by restricting attention to single-player games that didn't go through any public trial periods, we minimize the risk that time varying quality could drive both the player numbers and the reviews. We use price history to control for potential firm responses to changes in the review status. Classic panel methods are then sufficient to identify the causal effect of reviews on sales. We find that transitioning from "Mixed" to "Mostly Positive" review bin increases the inflow of new players (sales) by almost 5%, and that promotion to "Positive" further increases sales by 3%. The effect is strongest for games with a lot of reviews and for later stages of the game's life cycle, suggesting that consumers are cautios to place weight on review labels when the underlying amount of reviews is small or is subject to further development.

To address the concerns that there could still exist time-varying factors that are correlated with reviews and sales, we then turn to using a regression discontinuity approach for identification. By, essentially, comparing games that are several percentage points away from transitioning to and from the "Mixed", "Mostly Positive", and "Positive" bins, we find that the first two transitions increase sales by 7-8%, and that upgrading to "Positive" increases sales by 2.5%. However, we show that this approach, extensively used in the literature, could suffer from identification problems. We show that games several percentage points away from each other in terms of the review score, might have drastically different amounts of reviews, and that upgrading to a better review bin might require acquiring large amounts of new positive reviews. We suggest and use a different discontinuity design, based on the raw number of extra reviews that a games needs to upgrade to a better category. We find that promotion to the "Mostly Positive" bin improves sales by 4%, and promotion to "Positive"

improves sales by 3%—effects that are more similar to the ones obtained from panel methods.

We apply regression discontinuity techniques in a dynamic setting, when subjects can cross the cutoffs multiple times in the sample. To the best of our knowledge, rigorous econometric analysis of such settings does not exits. We suggest that ivestigating such settings could be a fruitful avenue for further research.

A What Fraction of Customers Leave Reviews?

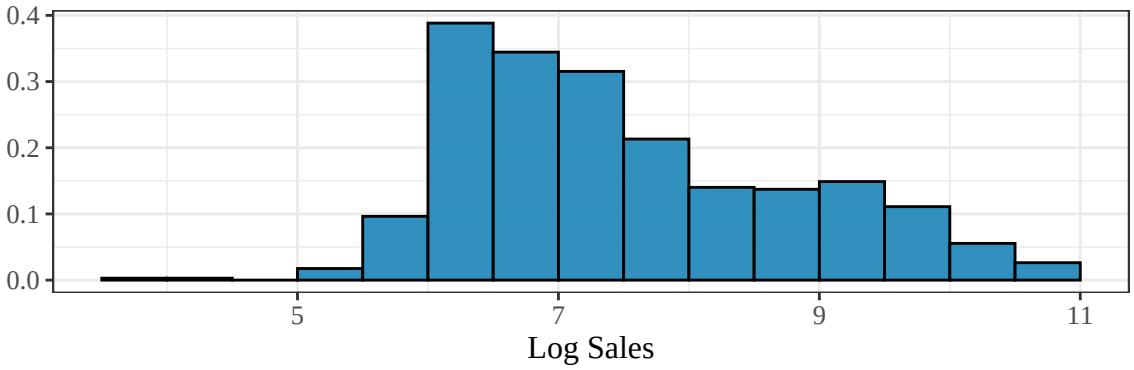

Figure 9: Estimated purchases in the first half a year of games' Life

The question of how many consumers leave a review is a question of separate interest that we can answer with our data. To our knowledge, there is not much evidence on what those numbers are in different markets. Clearly, the conversion ratio of customers to reviews is known exactly to the platform or the firm selling the good, but such data is not always available to researchers. We develop a methodology that allows us to estimate purchases from our high frequency usage data and to obtain an estimate of the conversion ratio, which could be extended to other settings when only data on usage is available. We also use spikes in player count upon an introduction of a new discount[9] as a more precise, yet local, estimate of purchases, and regress the inflow of reviews on that measure of purchases to provide a different estimate of the conversion ratio.

Our approach would work best if we had dayli data on the number of individual users instead of the maximum number of concurrent players. As players do not necessarily all play at the same time, the latter number is likely to be less than the former by an unknown factor. Thus, the exercise below is more interesting from the methodological perspective.

Empirical model (1) provides a clear way of separating continuing players from new

[9] see Table 5

players. The model implies that daily purchases could be obtained as

$$purchases_{it} := g_{it} - \psi g_{it-1} - day_{it} = \mu_i + \lambda_{it} + (\tau_t - day_{it}) + \epsilon_{it}, \tag{5}$$

where day_{it} are game-specific day of the week time effects. Equation (5) classifies any spikes in player activity that are not due to continuing players or within-week seasonality as new purchases. An alternative metric excludes the residual ϵ_{it} from the equation. Total purchases, which is a sum of daily purchases across the entire history of the game, doesn't depend on the definition, as the sum of residuals for a fitted regression that includes a constant is 0.

Table 11: Median Number of Reviews After a Discount

	Dependent variable:
	Review Count
Buyers	0.037^{***}
	(0.003)
Constant	-0.074
	(0.047)
Observations	5,090
Note:	*p<0.1; **p<0.05; ***p<0.01

Ideally, we would like to estimate (2) *individually* for every game in the sample, not differencing away the fixed effect term μ_i, which measures the average inflow of players absent the fluctuations in reviews and prices. This is a relatively involved computation, as time effects are shared among games. As estimation of purchases is not the main focus of the paper, instead we estimate the version of (2) without the shared time effects:

$$g_{it} = \psi_i g_{it-1} + \alpha_i disc_{it} + \beta_i score_{it} + \mu_i + day_{it} + \epsilon_{it} \tag{6}$$

This approach allows us to simply fit a separate regression to each game. We then obtain purchases using (5), and add up across the entire history of each game. Results for purchases in the first half a year are presented in Figure (9). Our sample is truncated from below, so

the purchases numbers are truncated as well. Since we fitted an individual regression to each game, we report median outcomes, since the averages are contaminated by the outliers.

Our estimates suggest that a median game has sold 1346 units in the first half a year of its life cycle, which looks reasonable, given that the majority of the titles in our sample are relatively small indie projects. Figure (10) presents the distribution of the conversion ratios over games in our sample. We find that the median conversion ratio is 8.36 percent.

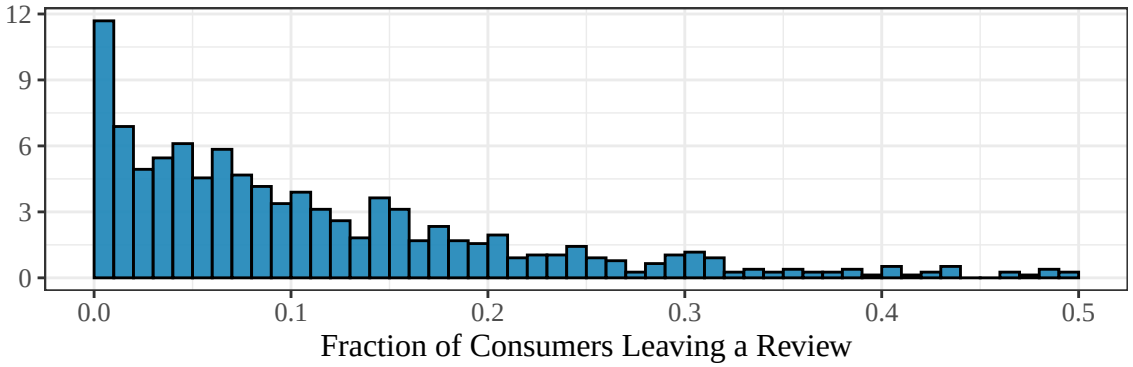

Figure 10: Estimated purhcases-to-reviews conversion ratios

Our estimate of the conversion ratio relies on the accuracy of our approach to estimating purchases. As a robustness exercise, we provide a related estimate using a different estimate of purchases. As mentioned before, discounts are usually followed by big spikes in player activity in the data, which provides us with a clean way to get an estimate of new purchases for the days when discounts are rolled out. Table (11) reports the results of a median regression of the number of new reviews on the day of the discount on the estimated purchases on the discount. We estimate that for a median game 3.68% of new buyers leave a review within one day. This number is consistent with the estimate obtained from the analysis of total purchases and reviews above, which calculates the propensity to leave a review when a user is presented with an arbitrarty amount of time to leave a review.

Akhmedov, Akhmed, and Ekaterina Zhuravskaya. 2004. "Opportunistic Political Cycles: Test in a Young Democracy Setting." *The Quarterly Journal of Economics* 119 (4): 1301–38. http://www.jstor.org/stable/25098719.

Anderson, Michael, and Jeremy Magruder. 2012. "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *The Economic Journal* 122 (563).

Anderson, T. W., and Cheng Hsiao. 1982. "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics* 18 (1): 47–82. https://doi.org/https://doi.org/10.1016/0304-4076(82)90095-1.

Association, Entertainment Software. 2019. "Breakdown of U.s. Computer and Video Game Sales from 2009 to 2017, by Delivery Format." *Statista*. https://www.statista.com/statistics/190225/digital-and-physical-game-sales-in-the-us-since-2009/.

Burtch, Gordon, Yili Hong, Ravi Bapna, and Vladas Griskevicius. 2018. "Stimulating Online Reviews by Combining Financial Incentives and Social Norms." *Management Science* 64 (5): 2065–82.

Chevalier, Judith, and Dina Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research* 43: 345–54.

Cliff, Edwards. 2013. "Valve Lines up Console Partners in Challenge to Microsoft, Sony." *Valve Lines up Console Partners in Challenge to Microsoft, Sony*. https://web.archive.org/web/20140914033025/http://www.bloomberg.com/news/2013-11-04/valve-lines-up-console-partners-in-challenge-to-microsoft-sony.html.

Dai, Weijia, Ginger Jin, Jungmin Lee, and Michael Luca. 2018. "Aggregation of Consumer Ratings: An Application to Yelp.com." *Quantitative Marketing Economics* 16: 289–339.

Documentation, Steamworks. n.d. "Discounting." https://partner.steamgames.com/doc/marketing/discounts.

Fradkin, Andrey, Elena Grewal, and David Holtz. 2019. "Reciprocity in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb." *Working Paper*.

Hollenbeck, Brett, Sridhar Moorthy, and David Proserpio. 2019. "Advertising Strategy in the Presence of Reviews: An Empirical Analysis." *Working Paper*.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35. https://doi.org/https://doi.org/10.1016/j.jeconom.2007.05.001.

Jian, Lian, Jeffrey MacKie-Mason, and Paul Resnic. 2010. "I Scratched Yours: The Prevalence of Reciprocation in Feedback Provision on eBay." *The Berkeley Journal of Economic Analysis & Policy* 10 (1).

Lee, David. 2008. "Randomized Experiments from Non-Random Selection in U.s. House Elections." *Journal of Econometrics* 142: 675–97.

Luca, Michael. 2016. "Reviews, Reputation and Revenue: The Case of Yelp.com." *Working Paper*.

Minotti, Mike. 2019. "NPD: U.S. Game Sales Hit a Record $43.4 Billion in 2018." *NPD: U.S. Game Sales Hit a Record $43.4 Billion in 2018*. https://venturebeat.com/2019/01/22/npd-u-s-game-sales-hit-a-record-43-4-billion-in-2018/.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–26. http://www.jstor.org/stable/1911408.

"Positive "Review Bombs"." 2019. https://steamcommunity.com/games/593110/announcements/detail/1621770561051427036.

Prause, Martin, and Jurgen Weigan. 2019. "The Shape of Things to Come." In *From Industrial Organization to Entrepreneurship: A Tribute to David B. Audretsch*, edited by Erik E. Lehmann and Max Keilbach, 99–120. Springer Nature.

"Steam Reviews Now in Beta." 2013. https://store.steampowered.com/news/11953/.

Zhu, Feng, and Xiaoquan Zhang. 2010. "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics." *Journal of Marketing* 74 (2): 133–48.