STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Academic year 2018–2019

The Changing Importance of Ratings for Sales:
Evidence from the Film Industry

Bendik Enstad (41049)

Abstract:

This thesis analyses the influencing effect that ratings have on revenue using a fixed effects framework. A panel data set of up to 2,025 movies running in US and Canadian cinemas with daily frequency is constructed. The period spans from 2010 to 2018. Critic ratings have a statistically and economically significant effect on sales, but clear evidence lacks for anonymous internet users to have a positive effect. More reviews by critics per se are also positive for sales. The effect of ratings is more important later in a movie's lifecycle but the analysis fails to detect a significant trend for ratings to generally strengthen over time. Furthermore, movies produced by the seven largest film studios in the US are more sensitive to average ratings than their smaller counterparts are. The revenue of drama movies are not more prone to ratings than that of other genres.

Keywords: Ratings, word-of-mouth, social learning, film

JEL: D12, D83, L82

Supervisor: Matilda Orth
Date submitted: 9 December 2018
Date examined: 8 January 2019
Discussant: Andreas Thürlimann
Examiner: Maria Perrotta Berlin

Acknowledgments:

I thank Matilda for being my supervisor and providing valuable feedback. I am also grateful to Maëlle, Øyvind, and Einar for reading through earlier versions of the thesis and highlighting areas for improvement. Remaining errors are entirely my own.

Table of Contents:

List of Figures:

List of Tables:

1 Introduction

Uncertainty shrouds many economic transactions. The quality of experience goods for instance, is difficult to know until the good is consumed and the cost incurred. A customer can never truly know ex ante whether a meal from restaurant "A" provides more satisfaction than that of restaurant "B", or which movie from the universe of all movies she is going to enjoy the most. The presence of uncertainty leads to more suboptimal choices than without it, as each purchase is essentially a lottery of utility. Nonetheless, an antidote to uncertainty is information, which might serve to guide our choices in such an environment. The last century has seen an accelerating boom in the spread of information through IT, which now enables us to access a vast depository of information unheard of in history at a hitherto unrivalled speed. This increased information access may thus have strengthened and may further yet strengthen the role of information in influencing economic actions. One such type of information is rating scores. Ratings have the property of reducing the multifaceted structure of a good down to a one-dimensional scale. Typically, this is in the form of number of stars, number of likes to dislikes, or a 1-10 scale. The information is thus in an easily digestible format with a low marginal cost. Hence, it is conceivable that ratings is a serious factor in explaining economic behaviour and thus worthy of academic attention.

The present thesis aims to shine a light on the relationship between ratings and sales. A panel data set of daily movie box office revenues with their associated average ratings is constructed. The data cover the period from 22 October 2010 up to and including 21 October 2018. There are two types of ratings, those from "critics" and those from anonymous "users". With "critic" is here understood a professional who specialises in assessing the quality of a movie and writing an associated review. A "user" on the other hand, is a non-professional who shares his or her opinion of a movie on an online forum. The thesis assesses both of these types of ratings on their ability to influence sales. The thesis then goes on to see how the effect of critic ratings may change over time and product. This includes tests to see whether ratings have a stronger influence later in a movie's lifecycle than earlier, and if there has been a wider trend of ratings strengthening in importance over the years. Furthermore, the paper presents tests as to whether movies from major film studios are more sensitive to changes in ratings than those of lesser-known studios are. Finally, as a response to an exploratory analysis done by Reinstein and Snyder (2005), the study evaluates whether revenues from "drama" movies are more prone to critic ratings than other genres.

The results indicate that critic ratings do indeed influence sales, but find no positive effect from user ratings more favourable than the average. On the contrary, the estimates do if anything point to a negative relationship between positive user ratings and revenue. A subsequent discussion of this puzzle and further examination reveals possible solutions; when controlling for the average critic rating, a higher user rating is associated with higher revenue for movies with sufficiently low critic ratings. The results provide evidence for movie revenue to be more sensitive to the average value of critic ratings later in the movie's lifecycle. The analysis detects no significant trend for ratings to strengthen in importance over the eight-year period. Movies from the biggest seven film studios are much more sensitive to changes in average ratings, but drama movies are not so compared to other genres.

The thesis contributes to both expanding a small literature as well as providing novel research. The relationship between ratings and sales has been studied before in restaurants (Cai, Chen, and Fang, 2009), books (Berger, Sorensen, and Rasmussen, 2010) and wine (Friberg and Grönqvist, 2012), all suggesting a positive relationship. And like this thesis, there exist studies of ratings focusing on the film industry (Reinstein and Snyder, 2005; Duan, Gu, and Whinston, 2008) which find no significant results, but these have comparatively small data sets. Reinstein and Snyder (2005) have 609 movie/time observations focusing on the effect of two prominent critics whilst Duan et al. (2008) have 71 movies in their sample studying anonymous users. In comparison, the present thesis employs a dataset of 2,025 movies for critics (1,754 for users) with more than 126,412 movie/day observations (120,710 for users) studying major US film critics. The thesis also explores new aspects of the relationship between ratings and sales. As far as I know, there is no study that empirically tests whether average ratings become more important over a product's lifecycle (although Friberg and Grönqvist (2012) study the persistence effect of reviews), whether a trend of ratings strengthening over time exists, or whether products of big establishments are more prone to ratings.

Besides previous studies on the film industry lacking material data-wise, there are other reasons to channel the focus onto this industry. The econometric design relies on time-variation of both ratings and revenue. One crucial assumption is that the underlying product does not change over time. Once a movie is on the market, the version is same yesterday, today, and forever.[1] This is in contrast to say, a restaurant or a hotel, which experience may be very reliant on time-varying conditions. For instance, the star chef may be temporarily unavailable or construction may take place in a hotel during a certain period, affecting both revenue and ratings without necessarily a causal link. Films are also typically a one-shot purchase, wherein learning the attributes from one self is less valuable. This stays in contrast to a repeat-purchase good such as wine, where testing something out for oneself can still provide useful information, regardless of whether the product is in taste or not. Learning from self and learning from others may be substitutes, and thus the lessened value in one may strengthen the demand for the other. Hence, learning from others through ratings may have more relevance in the film industry than in other sectors. Finally, data on box office receipts and ratings are publicly available at a daily frequency, making the study more transparent for validation and replication.

Academic interest aside, the study focuses on a pervading aspect of daily life where real money changes hands. A business would be better off knowing the extent to which ratings are worth being bothered by, and a policymaker may benefit in knowing how ratings can influence economic activity.

The paper proceeds as follows. Section 2 provides a background to give some context to the reader, and defines some key concepts. Section 3 reviews current literature both on social learning and on the relationship between reviews and sales. Section 4 formalises the motivation behind the study into testable hypotheses. Section 5 describes the data collection process and summarises key statistics. Section 6 lays out the econometric strategy that is used to test the hypotheses, and section

[1] Even though movies can come in the "Director's cut" or re-released with enhancements such as 3D, these versions are distinctly separated by title and identified as such.

7 provides the results. Section 8 runs through some robustness checks. Further discussion of the findings are presented in section 9, and section 10 reflects on the limitations and prospects for further research. Section 11 concludes.

2 Background

This section gives some context for the thesis and defines some central concepts. Subsection 2.1 explains the mechanism through which ratings can influence, and subsection 2.2 explains the concept of box office.

2.1 Ratings, reviews, and channel of influence

After someone has watched a movie, they might broadcast their opinion to the World Wide Web. The nature of this opinion can take several forms. They can for instance write a review. A review is in this thesis understood as a piece of text explaining the positives and negatives of an underlying good (in this case movie). It provides more nuance than a simple rating score, which is here understood the value reflecting the overall sentiment. This is for instance the "3.5/5" stars, the "37/100", or the "B-". Often a review is accompanied by a rating. This is particularly so for critics, who often write a review and then add a rating to quickly summarise the overall experience.

A rating score ought to reflect the extent to which someone likes or dislikes something. Since such a score is condensed into a one-dimensional scale, it cannot portray all the aspects of a product. It can however, give a snapshot of how good or bad a product is overall, considering which aspects are good and bad, weighted by importance. A guest of a hotel may agree that the assortment in the mini bar is excellent for instance, but may not care much about it anyway. Hence, for that guest the state of the minibar has no bearing upon the rating score he might give. Another guest might find it highly relevant and let it be decisive in the review. The aggregation of many individual ratings into an average score thus irons out the idiosyncratic dimension and gives an indication of quality weighted by what the average person finds important.

Average rating scores are common. For movies, the average rating is often used as a metric to gauge overall reception. Several metrics exist, but in the film industry three major ones are "IMDb"[2], "Rotten Tomatoes"[3], and "Metacritic"[4]. IMDb has an Alexa rank of 30 in the US and 54 globally as of 26 November 2018 (Alexa, 2018a), making it the largest of its kind. IMDb allows users of the site to post reviews and submit ratings, but the website also displays the "Metascore", a weighted average of the critic ratings given for a movie at Metacritic.[5] Major search engines such as Google report the averages of these three metrics if searching for a movie. Thus, if a consumer

[2] Website URL: https://www.imdb.com [last accessed 26 November 2018]
[3] Website URL: https://www.rottentomatoes.com [last accessed 26 November 2018]
[4] Website URL: https://www.metacritic.com [last accessed 26 November 2018]
[5] The actual weights themselves are a secret of Metacritic, and not publicly available.

is considering watching a certain film, she might research what others say about it and will quickly encounter these numbers.

Another way consumers might encounter ratings is by reading the independent reviews themselves. It is certainly possible that despite averages giving the general perception, consumers will find some critics more to their liking.

A change in average ratings therefore, might reflect influence of sales through two channels. The first, more direct route is by moving the actual average score that is widely reported. The second route is indirect through consumers observing the individual ratings themselves rather than the average. If an average rating increases for instance, then a new review has come out that is more positive than the average of previous ones. A higher rating than the average may thus be considered good for that particular movie. The scope of this thesis is not to decompose the influencing effect by average score and independent reviews, but rather seek to detect if ratings in general can influence sales, in addition to how any such effect may change over time and product.

2.2 Box office

After a movie is produced, it is then distributed to cinemas for display. The movie then runs for a given period before being taken off the screens. Although some movies are re-released years later, the primary market exists for a certain period only (weeks to months after a premiere). The movies may then be released for the home video market or streaming services. However, this thesis concerns itself with the market for movies in cinemas. The market thus differs from many others in that the good is limited for a certain time.

A "box office" is traditionally the location from which tickets to an event are sold. Tallying up the box office receipts from multiple venues across a country makes up the box office revenue as referred to in this thesis. This revenue is split among interested parties, primarily the exhibitors and distributors, but also producers (if separate from distributor) and even actors in certain instances. The exact split and contract vary, but commonly this can be a sliding percentage of revenues less the exhibitor's allowance (termed the "nut" in the industry), which includes house expenses, insurance, electricity, and mortgage payments (Vogel, 2007, p. 121). For blockbuster movies, a concrete example of such a revenue sharing contract may be that 70% of this net revenue goes to the distributor in the first week, and then reducing this percentage by ten percentage points every two weeks (ibid.).

3 Literature Review

Most of the theoretical literature on social learning and word-of-mouth axiomatically assumes that reviews influence the consumer. The focus is usually on the wider implications with respect to market outcomes, for instance whether social learning may lead to overconfidence in the information. Subsection 3.1 reviews some influential papers on theoretical models with social

learning as input. Subsection 3.2 focuses more on the empirical studies in the area, which more closely relates to this thesis. The empirical studies usually narrow down the scope in seeking to establish whether reviews and ratings have an effect on demand.

3.1 Social learning: theoretical foundations

The simplest models on social learning involve economic agents observing the purchasing/investing outcome of others. For instance, if a hungry customer stands at a street with two restaurants unbeknownst to him, he might take into account the number of current customers present at the two restaurants to infer quality through popularity. Banerjee (1992) constructs a model where individuals decide sequentially to invest in one asset amongst many with the setting that only one asset provides positive return. Prior to deciding, each player may receive an exogenous signal, and observes the decision of players acting before him to invest as well. In such a setting, utility-maximising rational agents would recognise that other players' signals are just as valid as their own, and as players cluster their decisions on one asset, the likelihood of that asset being the one giving positive return increases. If players early in the game receive signals that indicate a suboptimal choice is the optimal one, this could cascade in subsequent players herding on the inefficient outcome. Smallwood and Conlisk (1979) present a model where consumers cannot observe the intrinsic quality of a product (defined in terms of breakdown probability), but in the context of repeated purchases, can punish certain brands that breaks down. If the consumers rely too much on the market share (which they can observe) of the brands as indicator of quality, herding occurs and may in some cases lead to an inefficient outcome.

The aforementioned models consider one dimension of social learning, namely that of popularity. Returning to our hungry customer, consider instead that he can ask a sample of existing customers what they think of the restaurant. This adds another dimension of information. McFadden and Train (1996) present a model where heterogeneous consumers can buy a product repeatedly over three periods. They can buy in the first period at which point no experiential information exists, or they can wait to the second period and learn from others. The third period, which represents the rest of time, allows customers who bought in the second period and disliked the product to discontinue buying it. Adding learning from others delays the sales cycle as agents wait for others to try the product first, making their own decisions more informed. Still, there is always a segment that will not wait for others because they have strong priors that they will like the product. The model predicts that in the presence of learning from others, there is an asymmetry in the market against new products. Niche products that only appeal to a minority lose customers by the presence of word-of-mouth.

3.2 Reviews and sales: empirical findings

Overall, the empirical literature on reviews and sales generally finds that positive reviews and ratings have a positive effect on sales outside the film industry. However, for the film industry, evidence of a general effect is lacking. Studies which have used box office data and either critics or user data, conclude that there is no strong evidence of any effect from ratings.

Recognised as being the first to empirically study whether critics have an influencing effect on sales, Eliashberg and Shugan (1997) look at box office results of 56 movies and a respective aggregate critic score. The reasoning of the paper is that ratings can have both an influencing effect as well as a predictor effect, i.e. reflect financial success without affecting it. To the extent that critics influence, the authors argue that this effect should be strongest early on after release as word-of-mouth and other information is less available. The results document a significant correlation between sales and critic scores in the later life cycle of a movie (from fifth week onwards) but not in the preceding opening weeks. Hence, the authors interpret this as critics serving more as a predictor of sales rather than an influence, because of the reasoning that influence should be strongest in the early weekends and the data reveal the opposite. Nonetheless, this conclusion is tentative. An alternative explanation could well lie in consumer heterogeneity. Some consumers might factor in learning from others (critics or otherwise) heavier in the decision making process. This consumer group might also wait longer to see a movie as the longer time passes, the more learning from others can be extracted.

Subsequent papers with more robust methodologies produce mixed results. Reinstein and Snyder (2005) assess the effect by two prominent critics on box office revenue. Some of the movies are reviewed in the weekend of the premiere and may thus influence opening weekend revenue. Others are reviewed later, where influence on the opening weekend is impossible, as the information is not available at the premiere. The analysis relies on this to see whether a recommendation (thumb up or down) in the opening weekend is significant in explaining opening weekend revenue relative to a recommendation in the following period. When both critics give thumbs up in the opening weekend, this only has a marginally significant effect on excess opening weekend revenue and does not qualify as solid evidence that critic scores influence sales. The study goes on to break down movies by genre, and finds a positive effect for drama movies. However, this is done in an exploratory manner, wherein the sample is split up and regressions run separately, each evaluated at the same significance level. The result should therefore be taken with caution, as the probability of false positives increases significantly under such circumstances (see e.g. Simmons, Nelson, and Simonsohn, 2011 for a demonstration). A study also exists on the effect of user rating on film revenue. Duan et al. (2008) use panel data to assess whether users on the site "Yahoo! Movies" influence box office revenues through online word-of-mouth of 71 movies released between July 2003 and May 2004. The results indicate no significant effect of the ratings, but the number of posts made by users of a film positively contributed to sales.

Other papers studying reviews and sales generally report positively significant effects. Cai et al. (2009) conduct a randomised natural field experiment in a Chinese restaurant chain during October 2006. The design consists of two treatment groups. In one treatment group, tables are equipped with a plaque outlining the five most popular dishes, whereas in the other, tables are equipped with a plaque outlining selected dishes, without specifying popularity or recommendation. The latter treatment group is set up to distinguish the recommendation effect from that of the saliency effect (i.e. having certain dishes receive extra attention may in itself affect sales regardless of recommendation). The control group consists of tables without any additional information than the menu. The results point to no significant saliency effect, but estimates increased demand by

13-20% for the top five most popular dishes at the tables with the plaque specifying those dishes as most popular, significant at the 1% level.

Friberg and Grönqvist (2012) find critic reviews of wines in six large Swedish media houses to influence demand for those wines in Sweden. The effect is statistically as well as economically significant, with a positive review raising demand for a wine by 3.1% the same week using point estimates. The effect is also quite persistent, in that it is observed more than 20 weeks following a review, although the magnitude of the effect gradually withers away over time. A negative review has effectively zero impact. They also find that the effect of reviews is stronger among wines that are more expensive. By utilising data from the state monopoly on alcohol and having strict regulation on advertising as well as price margins, the results of the paper should be stripped of several confounding factors thus strengthening the robustness of the results.

Berger et al. (2010) find positive critic reviews to increase purchase likelihood of books, but interestingly, negative reviews may also increase this likelihood. Further examination suggests that if the author of a book is well known, negative reviews hurt sales, but if the author is new, a negative review is better than no review. This likely is the result of a negative review entering the consumption set of people for lesser-known products, whereas for well-known authors, the book is already in the consumption set. This "awareness effect" is already known to exist in the music industry (Hendricks and Sorensen, 2009) where artists who make a "hit" and become famous increase sales of all other albums (previous and forthcoming) than they otherwise would. It highlights the possibility that, insofar as it increase attention, controversies may pay off more the smaller a person/firm/product/idea is but actual rating and reputation becomes more valuable the larger any such unit is.

4 Hypotheses

This section formalises the motivation behind the thesis into testable hypotheses and corresponding null hypotheses. The null, which is the negation of the stated alternative hypothesis, must be rejected to count as evidence for the alternative hypothesis. This is done using standard significance tests such as the t-stat to compute whether the estimated effects are significantly different from zero.

Ratings provide an assessment of the quality of a movie, and as models on social learning postulate, yield relevant information to consumers. As previous empirical literature also suggests, favourable expert or critic ratings do have a positive effect (Berger et al., 2010; Friberg and Grönqvist, 2012). The first hypothesis H_1 and its corresponding null hypothesis H_1^0 is thereby:

(H_1) Higher critic ratings positively influence box office revenues.
(H_1^0) Higher critic ratings do not positively influence box office revenues of movies.

Although anonymous users on internet forums and sites may not have the same reputation as any famous critic, the ratings in aggregate still provide some level of information, and like critic scores,

a positive rating ought to reflect the positive sentiment towards a movie by fellow consumers. The consumer group has also been found influential in other studies (such as Cai et al., 2009). This leads to the postulation of the second hypothesis with its corresponding null:

(H_2) Higher user ratings positively influence box office revenues.
(H_2^0) Higher user ratings do not positively influence box office revenues of movies.

Controlling for ratings, number of reviews may also induce the awareness effect akin to Hendricks and Sorensen (2009) and Berger et al. (2010). For users, this is difficult to test due to endogeneity issues (since a user is a consumer, he has to watch the movie before writing a review). This is however tested for critic reviews:

(H_3) More critic reviews, ceteris paribus, lead to more box office revenues.
(H_3^0) More critic reviews, ceteris paribus, do not lead to more box office revenues.

Hypotheses 1-3 are similar to those of the papers mentioned in subsection 3.2, but are still of utility to test again on larger datasets. This thesis also tests new waters in the realm of ratings, in seeing how the effect of ratings may change over certain specifications. The next three hypotheses have to the best of my knowledge, never been tested before. Although two classes of ratings (critic and user) are available, the paper states the next hypotheses in terms of critic ratings only. Even though the logic behind the following hypotheses with the possible exception of H_7 should be independent of ratings type, having two hypotheses for each conjectured effect may lead to ambiguity of results and also a higher probability of false-positives as each effect is given two chances for detection. To avoid this, critic ratings are chosen as the primary material to use for the hypothesis testing as the quality of the data is arguably more solid (being both identifiable by name and reaching a larger audience). Still, having data on user rating easily enables the testing for the same hypotheses and shall be done. However, this serves more as a robustness check.

Different types of consumers may factor in the information of ratings differently in their purchasing decision. The model of McFadden and Train (1996) lets some consumers delay purchase to learn from others. As an alternative explanation to the results found in Eliashberg and Shugan (1997), certain consumers may wait longer and rely more on ratings. Other consumers, who have strong priors that they will like a movie, may watch the movie regardless of what others say, thus being less likely to wait for additional reviews and more word-of-mouth. Since the former consumer group may factor in ratings more heavily than the latter and are more likely to wait, ratings may have more influencing effect later in a movie's lifecycle. The next hypothesis formalises this belief:

(H_4) The average critic rating strengthens in importance the longer a movie stays on the market.
(H_4^0) The average critic rating does not strengthen in importance the longer a movie stays on the market.

Although there is to the best of my knowledge no other studies focusing on whether ratings generally strengthen over time, some inductive reasoning lends credence to this belief. As

mentioned in the Introduction, the last decades have seen an information explosion. Just in the last decade the rise of smartphones has mobilised the internet and the ability to consult the web for information on the go. Assume that (1) consumers use rating as an input in the decision making process (i.e. that H_1 is true), and (2) the progress in technology over the last years has made access to this information more widespread. Then, (3) more consumers have access to information that will be used as an input in their decision-making, strengthening that information's impact. Although the hypothesised trend is believed to span several decades, the available data for this study cover the last eight years. This should not be an exception, as the higher smartphone penetration in particular have mobilised internet making it more widespread. Thus, if someone decides to go to the cinemas while already out on the town and without easy access to a computer, they can still consult the internet through their phone to make a more informed decision. Hence, the next hypothesis states:

(H_5) Critic ratings have strengthened in importance over the last eight years.
(H_5^0) Critic ratings have not strengthened in importance over the last eight years.

Some movies are, for various reasons, better known than others are prior to release. This is partly due to larger advertising campaigns and more word-of-mouth. In accordance with the findings of Berger et al. (2010), more consumers will be aware of the product and the ratings may matter more. Although a movie might be well-known for a diverse set of reasons, big blockbuster movies which typically have a large production budget, more stars, and more resources to advertise, are also generally more well-known. Some film studios dominate much of the market. Previously, some studios went under the phrase "Big Eight" (Schatz, 1999, p. 47), but changing times have left some studios (such as MGM) to have declining influence. In the time-period considered in this study, seven studios consistently capture more of the domestic market than anyone else does. These are Walt Disney (previously Buena Vista), Warner Bros., Universal, 20th Century Fox, Sony / Columbia, Paramount, and Lionsgate. These seven studios have in the time period considered captured between 80-90% of the yearly market share among roughly 150 market participants considered. I will for the purposes of this thesis label this group as the "Big Seven". Hypothesis 6 thus states:

(H_6) Movies produced by the Big Seven are more sensitive to critic ratings than other movies.
(H_6^0) Movies produced by the Big Seven are not more sensitive to critic ratings than other movies.

Reinstein and Snyder (2005) find critic ratings to possibly affect drama movies more than other genres. Although the analysis is done in an exploratory manner, which lessens the meaning of the p-value, it is not unthinkable to be the case. If the consumer segment of drama movies factors in critic ratings more heavily, then this effect may arise. Reinstein and Snyder (2005) appeal to intuition for "art movies" to be influenced by critics, with "highbrow" consumers being overrepresented in the former category. Drama movies may arguably rely more on critic-sensitive aspects such as plot and acting to appeal to the audience, instead of special effects and explosions. Regardless, to test this prediction more formally, hypothesis 7 states:

(H_7) Critic ratings influence the revenue of drama movies more than other genres.
(H_7^0) Critic ratings do not influence the revenue of drama movies more than other genres.

5 Data

This section first describes the process that generates the dataset in the thesis, and then in subsection 5.2, summarises the nature of the data through key statistics and illustrations. Subsection 5.3 looks specifically at the variation of the independent variables, which is vital for the fixed effects analysis.

5.1 Construction of the data set

To construct the dataset, I compile data from two primary sources. Daily box office revenues for movies running in American and Canadian cinemas are reported by "Box Office Mojo"[6], which is the first data source used in this study. The website systematically tabulates observations with rows containing the title of the movie and the gross revenue earned in American and Canadian cinemas on a particular date, denoted in nominal USD.[7] The choice of US and Canada as geographical region for study is primarily because of data availability, as the website does not provide daily frequency data for other countries. Still, this market is currently the largest globally (Motion Picture Association of America, 2018), and is thus economically relevant. Box Office Mojo also has dedicated sections for the movies with additional information on certain characteristics. This includes genre and distributor, which hypotheses H_6 and H_7 relies on. The titles of the movies are subsequently used as an input in gleaning information about ratings, which is where the second data source comes in.

The second data source is Metacritic. Reviews and ratings on movies are scattered over many websites and media outlets, but Metacritic aggregates reviews from selected (mainly US) critics into dedicated sections on their website. This makes data harvesting a lot easier, and gives a diverse yet credible set of critics to study the effects on revenue. Reviews contain both a piece of text and a rating score, which differs in form depending on the media house that publishes the review. For instance, this can be number of stars (e.g. 2.5 stars out of 4) or an A-F letter grade. Such comparisons are not straightforward, but Metacritic normalises these rating scores so that they conform on a 0-100 point scale, where higher is more positive. Reviews that originally do not have a score is assigned one based upon Metacritic's impression of the review. Critics post reviews at different times. Some prior to the premiere, and some at various times after. This is crucial for the analysis as it produces time-variation in the average ratings in the period wherein the movie runs in the theatres. For the time-series of box-office revenues obtained from Box Office Mojo, each eligible movie/date observation is appended a data point of the average rating from critics reviewing the respective movie. This is a simple mean of all ratings from reviews published up to

[6] Website URL: https://www.boxofficemojo.com [last accessed 21 November 2018]
[7] Conversion to real revenue is not necessary. The econometric design (presented in section 6. Econometric Strategy) controls for date dummies and a time trend. Inflation is thus loaded onto these terms and accounted for.

and including the date of observation. Also included is the number of reviews from which the ratings are drawn In total, 52,277 unique critic reviews from 90 media establishments constitute the basis for these calculations. The names of these media houses can be found in subsection A.1 in the appendix.

Although a plethora of internet forums where anonymous users can share their opinion of a movie exists, Metacritic also has such a function. Since Metacritic is the source of critic reviews, the data collection process is readily implemented for user reviews as well. Metacritic has an Alexa rank (measure of most popular websites on the internet) of 587 in the US and 1,295 worldwide as of 26 November 2018 (Alexa, 2018b). Thus, it is not the most popular online forum for users discussing movies. That title goes to IMDb, which concurrently has an Alexa rank of 29 in the US and 53 worldwide (Alexa, 2018a). In order to detect any influencing effect that users have on revenue, it would probably have highest probability of success studying IMDb data because of the traffic. However, technical hindrances in the data collection process makes this less viable.[8] Hence, Metacritic constitutes the source for user reviews as well as critic reviews. The user reviews contain a score from 0-10, again higher being positive. For the purposes of this thesis however, this score is multiplied by 10 so that it is commensurate with the 0-100 scale used for critic scores making comparisons of any effect direct. The calculation for user ratings is akin to that for critics; all user reviews up to a given date forms the basis of the average score on that date. Metacritic allows users to both post a review and to assign a rating without posting a review. These non-review ratings are not time-stamped on the site, so is not included in the calculation of the average rating. The method relies on time-variation of the data. Therefore, knowing the time at which a rating has been submitted is paramount. In total, the calculations stem from 105,036 unique user reviews.

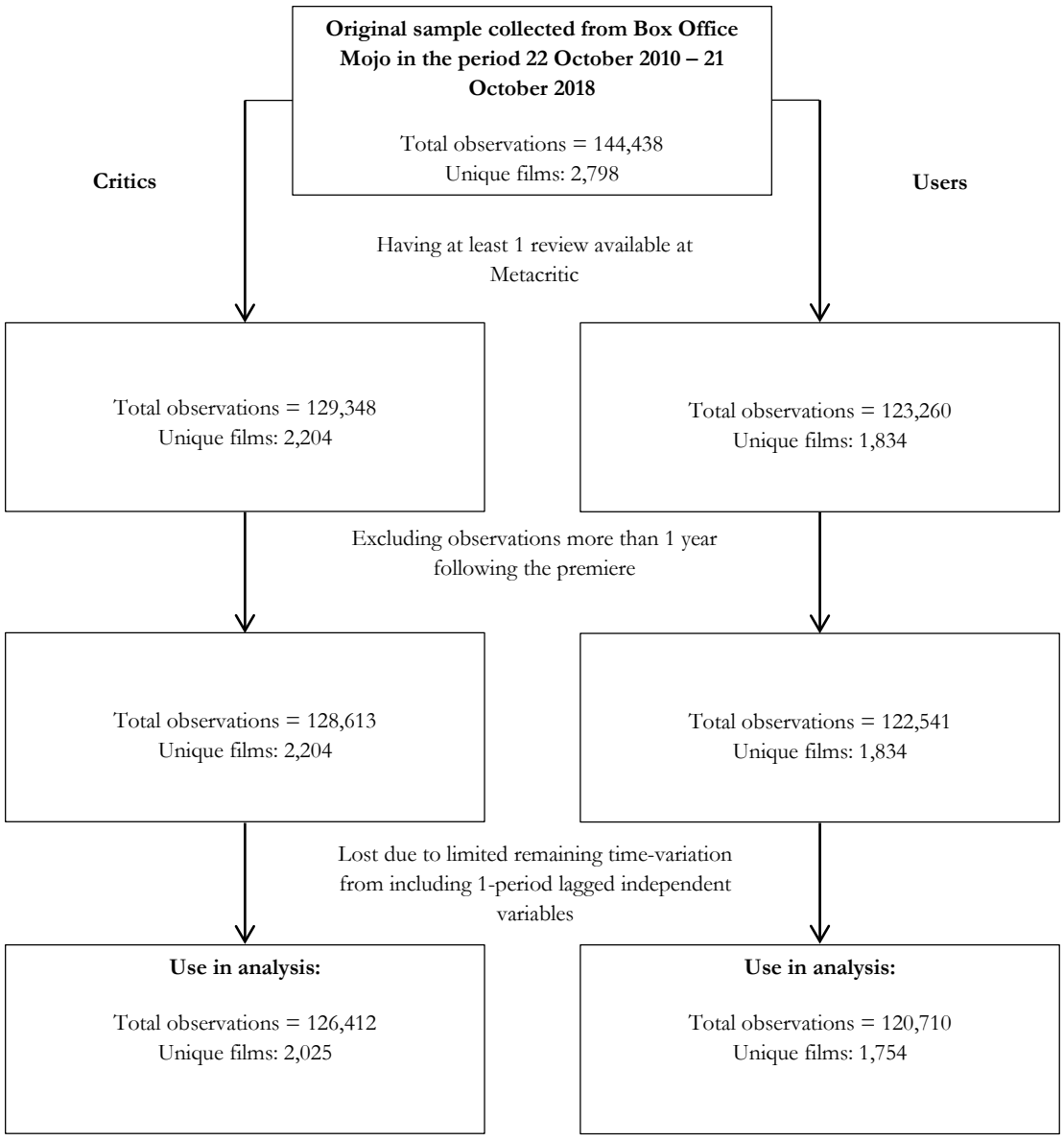
Some of the entries in the data from Box Office Mojo are excluded in the dataset. First, critic reviews for movies prior to 22 October 2010 are not time-stamped on Metacritic. This makes the average rating on a given day impossible to compute without further information from external sources. Hence, the time period studied is restricted from 22 October 2010 up to and including 21 October 2018. This marks exactly eight years' worth of data. Second, some of the titles on Box Office Mojo do not have review sections on Metacritic. This includes movies with tiny box office revenue, movies that are foreign in origin (typically Indian), non-movie entertainment such as certain boxing matches that were shown in cinemas, and re-releases. Some observations are for many years after the movie first premiered, the maximum being 2,816 days after the premiere ("Nostalgia for the light" on 29 September 2018). As days since premiere is a control variable, such extreme outliers will likely skew the results and a cut-off for observations greater than 1 year (365 days) is thus added. The sensitivity of this decision's impact on the conclusions is tested for in subsection 8.2.

The econometric strategy outlined in section 6 uses a 1-day lag of some independent variables, which consequently drops one observation for each movie. As some movies have no time-variation after this observation is dropped, the final sample used in the analysis is further reduced. The final

[8] Specifically, IMDb has the functionality of the reviews written in the programming language JavaScript, which makes the reviews elusive in the source code for the website. To obtain all the reviews requires literally clicking a button multiple times until the reviews have loaded. This is cumbersome, time-consuming, and prone to error for both a web-crawler (a program that systematically downloads data from the internet) and a human.

result of this data harvesting process and purging thereof yields a dataset consisting of up to 2,025 movies for critics and 1,754 for users. A summary of this process is outlined in Figure (1).

Figure (1) – Schematic of data collection process:



5.2 Summary statistics

Table (1) contains summary statistics of the dataset. The average daily revenue for a movie is USD 627,099, but there is substantial variation as the standard deviation is more than four times as great at USD 2,532,500. This is not surprising; revenues tend to be concentrated near the premiere and on weekends, and some blockbuster titles capture a lot of market share. The highest observed

revenue was the premiere of "Star Wars: The Force Awakens", which on that day earned USD 119 million in domestic (read US and Canada) revenue (over the course of its lifetime it earned USD 937 million domestically). Six observations show USD 4 in daily revenue. These are all months after the premiere and for comparatively small titles. Average critic rating has a mean of 61.23 with the median slightly higher at 62.14. For users, the mean of the average ratings is higher than that of critics, at 65.40, with median 67.78. The standard deviation is also somewhat higher for users (17.23) than for critics (15.82), so one might say that users are on average slightly more generous with their ratings but give more extreme assessments. Figure (2) illustrates the distribution of average ratings in the dataset. The average movie runs for 47.92 days in the theatres, thus providing a decent number of time observations for each movie. The average number of critics who review a movie is 30.23 (median 32), and for users the average is 45.82 (median 16), but there is considerable variation. Particularly so for users, which number of reviews ranges from one to 3,393 ("Star Wars: The Last Jedi").

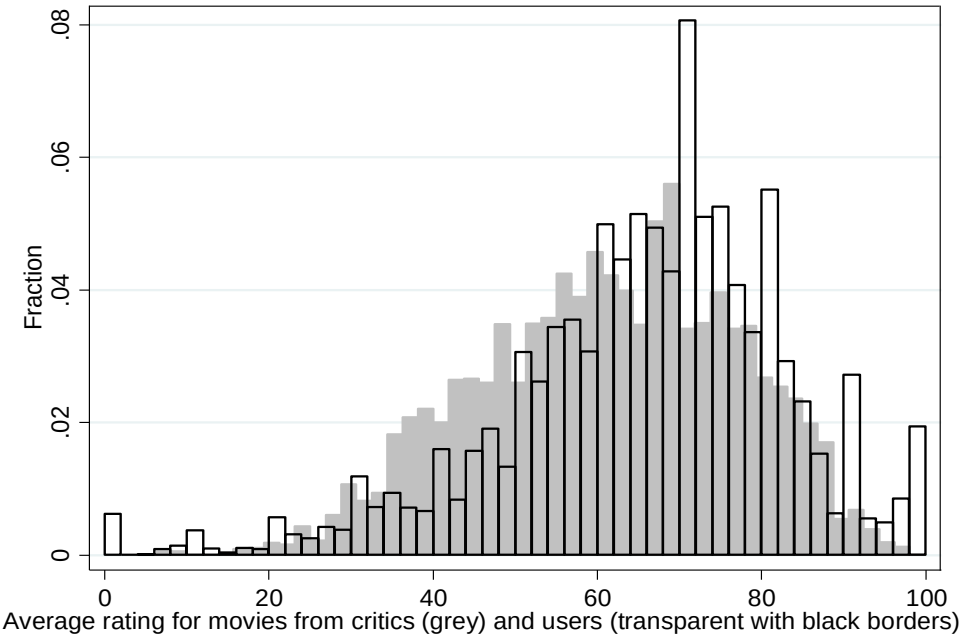
Table (1) – Summary statistics:

Variable	N	Mean	SD	Min	25%	Median	75%	Max
Revenue (USD)	132,809	627,099	2,532,500	4	6,524	37,450	276,836	$1.19 \cdot 10^8$
Average critic rating (0-100)	128,613	61.23	15.82	4.43	49.73	62.14	73.39	100
Average user rating (0-100)	122,541	65.40	17.23	0	56.45	67.78	76.73	100
Average number of critic reviews	128,613	30.23	13.12	1	21	32	40	60
Average number of user reviews	122,541	45.82	132.46	1	6	16	43	3,393
Days since premiere	132,809	47.92	41.16	0	17	39	68	364

Note: Each observation constitutes a movie on a day. Not all observations have both a critic and a user review, leading to a smaller set of observations for average critic and average user rating than for revenue and days since premiere. 2,204 movies have critic rating information, and 1,834 movies have user rating information. 2,315 movies have either critic or user rating information.

Source: Author's collection and rendering of data from Box Office Mojo (revenue) and Metacritic (ratings and reviews).

Figure (2) – Distribution of ratings:



Note: The figure shows the distribution of the average critic and user ratings (on a scale from 0-100) for all movie/date observations in the dataset.
Source: Author's collection and rendering of Metacritic data

Table (2) presents the correlation matrix for the main variables used in the analysis. Revenue is transformed to its logarithmic form since that is what the analysis use. Several surprising results are worth noting. Most strikingly is that log(revenue) and average critic score have a very low correlation, and a negative of that, whilst the correlation between log(revenue) and average user score is virtually zero. This is, on the face of it, completely contrary to what one would expect. This suggests that ratings have a weak association with revenue. The correlation coefficient assumes a linear relationship and includes all movie/date observations. It thus also includes days where an otherwise successful film receives lower revenue (further from the premiere). To further investigate this, Figure (3) provides a scatter plot of the log of cumulative revenue on the y-axis and average ratings from all reviews on the x-axis, both observations from the last date of observation (i.e. when all revenues and ratings have been tallied up). Yet the figure hardly reveals a strong relationship at all, linear or otherwise. Despite this, it is still possible that ratings have an effect given a certain movie's characteristics, as ratings may correlate with some other unexplained variables. However, any such correlation would if anything be expected positive. Even if it does not influence sales, a rating ought to reflect quality, which also should generate sales. The dataset includes large and small movies alike. If one considers the movies of the biggest seven movie studios only, the correlation coefficient turns to 0.0784 for critic and 0.0688 for user, both statistically significant from zero at the 1% level. This might give an indication that ratings matter more for bigger movie studios, but the correlation is still low.

Table (2) – Correlation matrix

	log(revenue)	Average critic rating	Average user rating	Number of critic reviews	Number of user reviews	Days since premiere	Date [a]	Big Seven [b]	Drama [c]
log(revenue)	1								
Average critic rating	−0.0329***	1							
Average user rating	0.0005	0.4894***	1						
Number of critic reviews	0.3924***	0.3853***	0.1046***	1					
Number of user reviews	0.1280***	0.1542***	0.0194***	0.3138***	1				
Days since premiere	−0.3445***	0.2037***	0.1033***	0.1924***	0.1465***	1			
Date [a]	−0.0639***	0.0942***	−0.0595***	0.0334***	0.0498***	−0.0411***	1		
Big Seven [b]	0.3995***	−0.0929***	−0.0788***	0.2828***	0.2059***	0.0352***	−0.0553***	1	
Drama [c]	−0.1140***	0.1928***	0.1082***	0.0417***	−0.1051***	−0.0154***	0.0364***	−0.2208***	1

Note: the table outlines the correlation between any two variables used in the dataset. Both average critic and user rating are on a 0-100 scale.
***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively. Observations are movie/day and total 119,062. Average ratings are drawn from all reviews available up until the day of observation and reported on Metacritic.

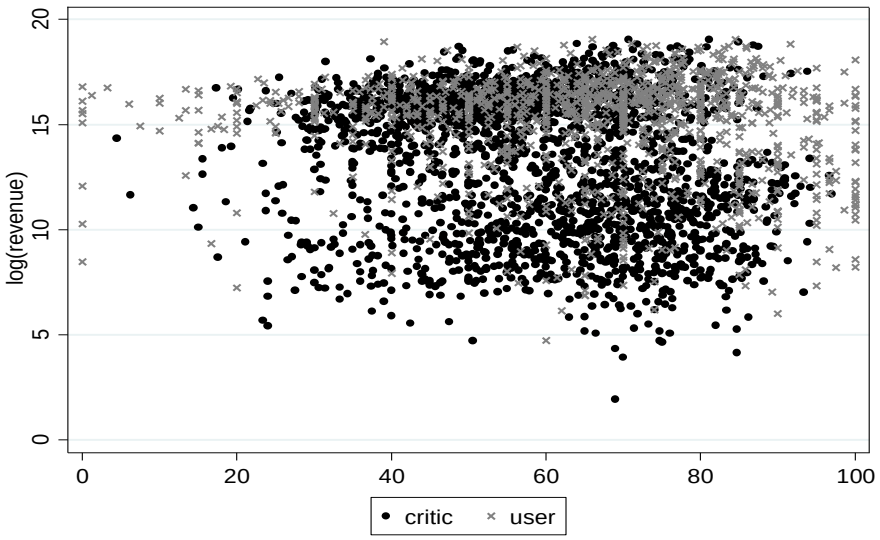
[a] Date is defined as 0 on 22 October 2010, and increases by 1 for each following day.

[b] Big Seven refers to whether a movie is distributed by one of the seven largest film studios in America (=1) or not (=0).

[c] Drama is a dummy variable indicating whether the genre is drama (=1) or not (=0).

Source: Author's collection and rendering of data from Box Office Mojo (revenue, date, Big Seven, drama) and Metacritic (ratings and reviews).

Figure (3) – Scatter plot of cumulative log(revenue) and average rating scores:



Note: The figure shows a scatter plot between log(revenue) for a movie on the y-axis and average rating score on the x-axis. The log(revenue) is not daily, but summed over the entire period a movie runs in the cinemas. Likewise, the ratings score are the average of all available ratings given to any movie. A black dot represents an observation for an average critic rating for a movie on a given day and a grey cross for user.

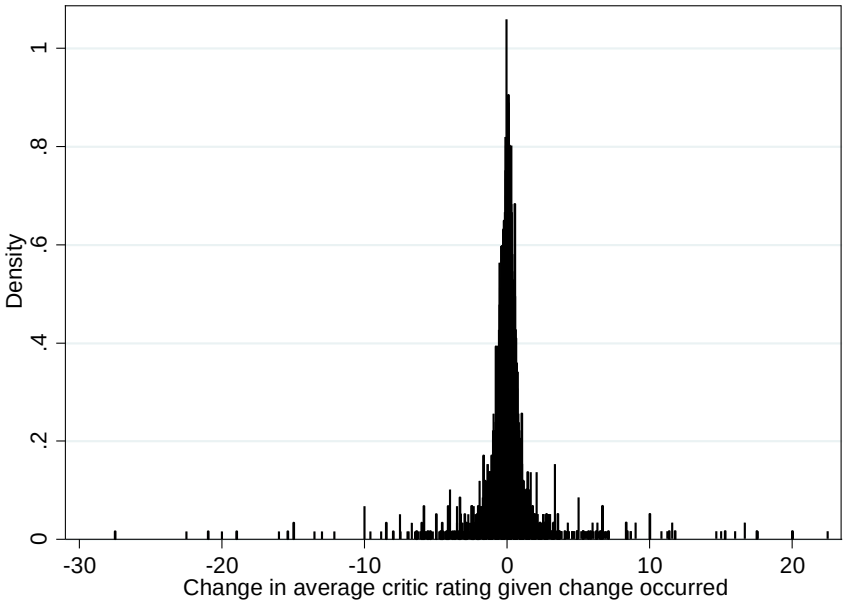
Source: Author's collection and rendering of data from Box Office Mojo (revenue) and Metacritic (ratings)

The fact that date has a negative relationship with log(revenue) might also have caught some by surprise. One would if anything expect inflation and economic growth to have increased revenue over time. The result can however be explained by more movies on the market, which in turn lowers the average market share of a movie. Aggregating the daily revenue of all movies reveals a positive relationship (correlation coefficient of 0.1655) between revenue and date as expected. Less surprising is that revenue is inversely linked to days since premiere. Higher grossing movies also correlates with number of reviews. The variable for days since premiere is positively correlated with both average critic and user scores. A possible explanation is that more successful movies would run longer in cinemas.

5.3 Variation of the independent variables

As the study relies on time variation, it is worth looking at the nature of the variation in the independent variables used in the analysis, namely average critic and user ratings, as well as number of reviews. Figure (4) shows the distribution of the change given a change occurred for critics, and Figure (5) does so for users. The distributions are centred around zero and is symmetrical. The change is on average -0.0177 (SD = 1.6990) for critics and -0.1614 (SD = 4.3172) for users. Thus, there is no severe systematic material bias for ratings to go either up or down consistently over time. A lot of variation is helpful in the estimation of the fixed effects regressions, but overall there is little variation in average ratings over time. Crucially however, there is still variation.

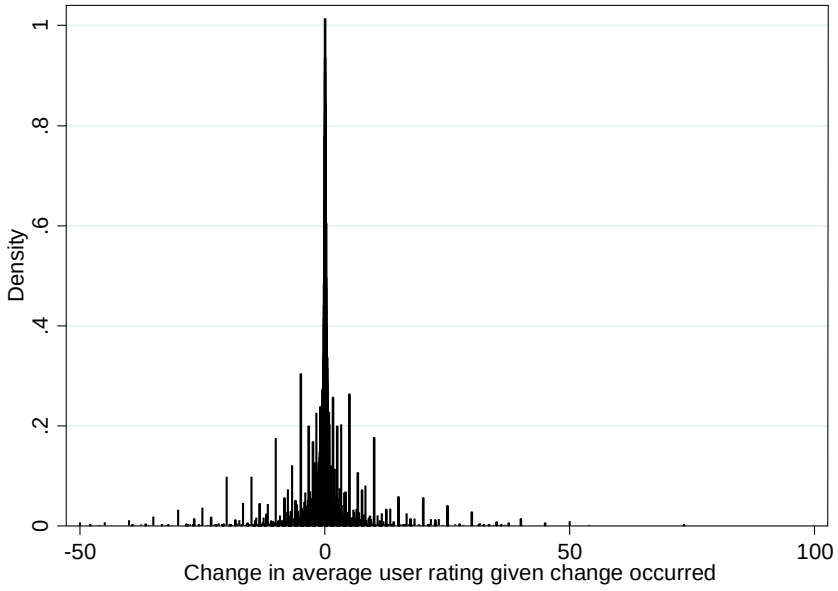
Figure (4) – Distribution of changes in average critic ratings:



Note: The figure shows the distribution of how much the average critic ratings changed from one day to another excluding days were no change occurred.

Source: Author's collection and rendering of Metacritic data.

Figure (5) – Distribution of changes in average user ratings:

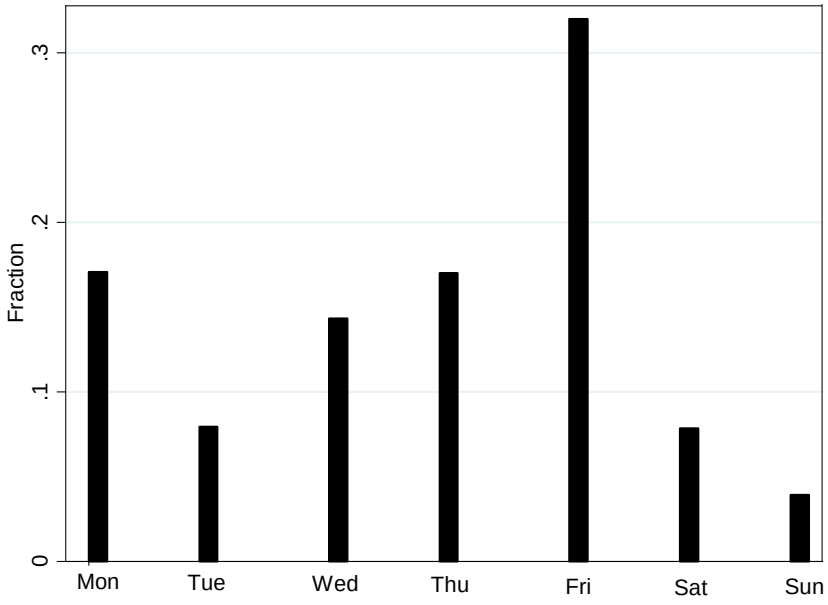


Note: The figure shows the distribution of how much the average user ratings changed from one day to another excluding days were no change occurred.

Source: Author's collection and rendering of Metacritic data.

The aggregate number of reviews is by nature either increasing or unchanged throughout the time-series for each movie as it is cumulative in form. Some reviews are posted prior to the premiere and some after. The average number of critic reviews on the premiere is 15.94, whereas the average number of critic reviews in the time following the premiere is 27.21. It is the reviews following the premiere that is instrumental in affecting the results of the analysis. If the timing of reviews is endogenous, then this could be of concern for hypothesis H_3, which predicts an increase in sales from the number of reviews (holding the rating of the review fixed). One way this could occur is if critics only post on certain days of the week. For instance, suppose reviews only came out on Friday, then higher weekend revenue may bias the estimate. Figure (6) shows the distribution of when the number of reviews changes by day. Most changes occur on Friday, but there are decent number of observations for all weekdays.

Figure (6) – Reviews by publishing day:



Note: The figure shows the daily distribution of when the number of reviews changed, i.e. when a new review was released.

Source: Author's collection and rendering of Metacritic data.

6 Econometric Strategy

The econometric strategy outlined in this section serves to evaluate the validity of the hypotheses stated in section 3. The approach exploits the time-varying nature of the data. Subsection 6.1 is the first step in describing how to assess the influencing effect of critic ratings (H_1), user ratings (H_2) and number of reviews (H_3) on sales. Subsection 6.2 then looks at how to evaluate the effect changes of ratings over time (H_4 and H_5), and subsection 6.3 outlines how to determine whether there is an additional effect on ratings from Big Seven movies (H_6) and drama movies (H_7). Subsection 6.4 addresses concerns on statistical inference with respect to heteroskedasticity and

autocorrelation. The rationale for using fixed effects instead of first differencing or random effects is presented in subsection 6.5.

6.1 The effect of ratings

The financial success of a movie running in the theatres is compounded by a diverse set of underlying explanations. Production budget, director, actors, and genre are all examples of observable variables, but the success is also susceptible to unobservable variables such as changing tastes, consumers' perception of a trailer or other form of advertisement, and offline word of mouth. To estimate the effect of ratings on a movie in this environment using OLS would be naïve, since ratings are likely correlated with many of the unobservable variables. Hence, a correlation between ratings and revenue is not sufficient to ascertain a causal influence of the former onto the latter.

To overcome these obstacles, one can exploit the time-varying nature of the data obtained in this study. Different ratings are published at different times, leading to a movie having different average ratings during a movie's life in the theatres. Since the movie with all its inert properties remains the same throughout this period, one can focus in on how the time-variation in ratings correlates with the time-variation in revenue. The first model to be estimated is:

$$\text{(M1)} \quad \log(revenue_{i,t}) = \beta_1 rating_{i,t-1} + \beta_2 Nreviews_{i,t-1} + \beta_3 sinceprem_{i,t} + \alpha_i + \delta_t + u_{i,t},$$

where $\log(revenue_{i,t})$ is the natural log of box office revenue for movie i on day t. The logarithm is used instead of actual revenue to capture the percentage increase in revenue rather than the absolute increase, thus permitting rating to have a proportional impact on small and large movies alike. $rating_{i,t-1}$ is the average rating, delayed by one day and stemming either from critics or users – depending on the specification, and $Nreviews_{i,t-1}$ is the number of reviews from which the ratings are calculated. The reason for the one-day time delay in $rating_{i,t-1}$ and $Nreviews_{i,t-1}$ is to allow the information to reach the market. $sinceprem_{i,t}$ refers to days since premiere and thus controls for the trend of depreciating revenue following the premiere. δ_t are date fixed effects, and allow each special date to have its unique influence on revenue. For instance, Saturdays and national holidays are likely to have higher revenue. This also controls for the Friday bias for number of reviews as illustrated in Figure (6) in subsection 5.3. α_i are movie fixed effects, and encompass all the unique traits associated with a movie. This includes the plot, the actors starred, the director, the trailer, and a plethora of other characteristics. $u_{i,t}$ is the unexplained error term, and the $\beta's$ are coefficients to be estimated.

For hypotheses (H_1) and (H_2), β_1 should be positive in (M1) when considering critic and user ratings respectively. For hypothesis (H_3), β_2 should also be positive. For users, the time at which reviews are posted is likely to be less exogenous than for critics. Users (who are consumers) should have to watch the movie before reviewing. This can induce a negative bias in β_2, as days where a movie have a lot of viewers can also produce more reviews, which can thus bias the estimate of β_2 downwards.

A natural next worry is whether the time-variation in the ratings is correlated with the time-variation of unobservable characteristics that are not controlled for. Although one can rarely completely rule out any such concern with ironclad certainty, an argument can be made for this to be of limited concern. A critic is still evaluating the same movie as those before her and if serious, should not be influenced by factors outside the realm of the movie's attributes, which are fixed over time. One can imagine each critic to partly be influenced by a common set of factors influencing all critics, as well as some idiosyncratic component local to the critic. These idiosyncratic differences drive the variation in the average rating scores. The analysis essentially assumes that these idiosyncratic components are exogenous. As the movie and thus the quality is unchanged over time, and a critic is supposed to rate the movie itself, this assumption seems sound.

Even if a scandal erupts surrounding an actor related to a movie for instance, this might influence sales, but still does not affect the quality of the movie. It is of course possible that the human psyche could sway a critic in this regard, but the gravitas of the rating will be on the actual movie anyway. In addition, such a scandal would have to erupt in the relatively short timeframe during a movie's run in the cinemas. When looking at scores given by users, this might be of greater concern; organised efforts to spam bad reviews may happen if a group dislikes a decision taken by the company for instance. Still, these events have to occur during the time wherein the movies are on the cinemas, and are presumably rare.

A more pressing issue might be advertising, which may change over time. It is not unrealistic to imagine advertising playing together with changing ratings. In fact, Elberse and Anand (2007) find evidence of ratings interacting positively with advertising expenditure in movies. Nonetheless, institutional barriers and industrial practices restrict the flexibility in adjusting advertising to changing ratings, even in the weeks prior to the movie's premiere. Industry executives report in interviews that the advertisement strategy is laid out months prior to release and the lion's share of advertising expenditures are spent ahead of the premiere (ibid.). Television advertisements, which constitutes the vast majority of the budget, typically does not allow purchased time slots to be unsold (Sissors and Baron, 2010, p. 349). Hence, the leeway for changing advertising in light of new reviews is small.

Even though prices may vary from cinema to cinema, or in price discrimination between customers (e.g. discount for elderly or children), US cinemas have exhibited a uniform pricing strategy since the 1970s and ticket prices are not varying with patterns of demand (Orbach and Einav, 2007). This study is mostly concerned with the overall impact ratings have on sales, but with limited variation in prices, the interpretation as of changes in demand is more direct. As the marginal cost of an extra customer seeing a movie is zero for the distributor, and also practically zero for the exhibitor, the variation in mark-ups follows that of price.

6.2 The time-changing importance of ratings

Two of the hypotheses in this thesis relate to how ratings may change in importance over time. The first is that ratings for a given movie increase in importance the longer time has passed since the premiere (H_4). The second is that the ratings have generally become more important over the

last 8 years (H_5). To evaluate these hypotheses, one can make use of interaction variables. For the days since premiere hypothesis, the effect of ratings should be higher the longer time has passed since the premiere. Hence, if ratings have more effect later in the lifecycle, the interaction term should have a positive coefficient. Adding this interaction to (M1) gives:

$$\text{(M2)} \quad \log(revenue_{i,t}) = \beta_1 rating_{i,t-1} + \beta_2 Nreviews_{i,t-1} + \beta_3 sinceprem_{i,t} + \beta_4 rating_{i,t-1} \times sinceprem_{i,t} + \alpha_i + \delta_t + u_{i,t}.$$

The coefficient of interest in (M2) is β_4, which measures how the effect of ratings changes as days since premiere passes. Hypothesis (H_4) predicts $\beta_4 > 0$, and the null $\beta_4 \leq 0$.

To test whether critic ratings have become more important over time (considering the last 8 years), an interaction variable between ratings and the date variable is constructed. Since a fixed effects regression is used, this interaction term cannot simply be added to equation (M1) however, since within each movie it will be highly correlated with the $rating_{i,t-1} \times sinceprem_{i,t}$ term and could thus produce biased results if there is indeed an interaction effect between rating and days since premiere. Hence, the $rating_{i,t-1} \times t$ term has to be added to equation (M2) instead, so that $rating_{i,t-1} \times sinceprem_{i,t}$ is controlled for:

$$\text{(M3)} \quad \log(revenue_{i,t}) = \beta_1 rating_{i,t-1} + \beta_2 Nreviews_{i,t-1} + \beta_3 sinceprem_{i,t} + \beta_4 rating_{i,t-1} \times sinceprem_{i,t} + \beta_5 rating_{i,t-1} \times t + \alpha_i + \delta_t + u_{i,t}.$$

If there is a trend of ratings increasing in importance over time, then $\beta_5 > 0$, as hypothesis (H_5) states. For it to count as evidence, its corresponding null, which translates to $\beta_5 \leq 0$ would have to be rejected.

6.3 The category-changing importance of ratings

Hypothesis (H_6) is motivated by the possibility that ratings ought to have higher importance the better known a product is. To find out whether the movies produced by the Big Seven are more prone to fluctuations in ratings, an interaction variable between rating (delayed by one day) and a dummy equal to one if a Big Seven studio distributes a movie is generated. This is then added to equation (M1):

$$\text{(M4)} \quad \log(revenue_{i,t}) = \beta_1 rating_{i,t-1} + \beta_2 Nreviews_{i,t-1} + \beta_3 sinceprem_{i,t} + \beta_6 rating_{i,t-1} \times big7_i + \alpha_i + \delta_t + u_{i,t}.$$

If ratings have a relatively more influencing effect on movies produced by the Big Seven, then coefficient $\beta_6 > 0$. The null to be rejected is $\beta_6 \leq 0$.

The final hypothesis (H_7) is that drama movies are more prone to the influence of ratings than other types of movies. An interaction between rating and a dummy equal one if the movie belongs to the drama category is constructed, and (M1) modifies to:

$$\text{(M5)} \quad \log(revenue_{i,t}) = \beta_1 rating_{i,t-1} + \beta_2 Nreviews_{i,t-1} + \beta_3 sinceprem_{i,t} + \beta_7 rating_{i,t-1} \times drama_i + \alpha_i + \delta_t + u_{i,t}.$$

If the effect of rating is reinforced by belonging to the drama category, then $\beta_7 > 0$ in accordance with hypothesis (H_7). The null (H_7^0) translates to $\beta_7 \leq 0$ in this case.

6.4 Heteroskedasticity and autocorrelation

In order to make proper statistical inference in models M1-M5 and evaluate the significance of the coefficients, the estimation of the standard errors must be consistent. This would fail in the presence of heteroskedasticity or autocorrelation of the errors. In terms of autocorrelation, the fixed effects procedure ensures that the serial correlation tends to zero as the number of time observations for each group grows (Woolridge, 2010, p. 270), thus rendering this concern low for the present dataset as the average group contains 48 time periods (see Table (1) in section 5). Heteroskedasticity of the errors on the other hand would result in inconsistent estimates of the standard errors, which could end in faulty conclusion for the hypotheses. Presence of heteroskedasticity in a fixed effects regression can be tested for using a modified Wald test (Greene, 2000, p. 598). If the null hypothesis that there is no heteroskedasticity is rejected, then one can cluster the standard errors at the level of movies to obtain robust standard errors that are consistent and can thus be used to properly evaluate the significance of the results (Woolridge, 2010, p. 271-272). As it turns out, for all of the following results in the subsequent section, the modified Wald test decisively rejects the null hypothesis of homoskedasticity of the errors. Thus, the standard errors reported for fixed effects regressions throughout this thesis are clustered at the level of movies and are robust.

6.5 Rationale for fixed effects versus alternatives

When dealing with panel data, fixed effects are not the only way to conduct analysis. Alternatives include first differencing and random effects models. Instead of time demeaning the data for each panel, first differencing subtracts the previous period's value from the current one for each variable employed in the regression. A first difference specification of (M1) is:

$$\text{(FD1)} \quad \Delta\log(revenue_{i,t}) = \beta_1 \Delta rating_{i,t-1} + \beta_2 \Delta Nreviews_{i,t-1} + \delta_t + \Delta u_{i,t},$$

where Δ indicates the change from the corresponding observation the previous day, and the rest of the notation is similar to that of (M1). $sinceprem_{i,t}$ is not included since it changes by one for each day making a differencing consists entirely of observations equal one. If rating increases on day $t-1$ from day $t-2$, model (FD1) estimates how revenue changes on day t from day $t-1$ through the coefficient β_1. Thus, if a portion of the increasing rating manifests itself on a later day, (FD1) fails to detect this. Instead, it only captures the immediate effect. The fixed effects specification however, such as (M1), regress log(revenue) on rating. Thus, time-varying levels rather than immediate differences are captured. If average rating decreases for instance, but does not

change for several days, the fixed effects specification treats each observation where rating is lower equally. As this is common (most days sees no change at all in average critic ratings), this speaks for using the fixed effects approach.

Still, in the presence of non-stationarity, a time-series can produce spurious regressions (Granger and Newbold, 1974). First-differencing can solve this and may therefore be preferred. To test for a unit root process in panel data, a modified Fisher test is recommended by Maddala and Wu (1999). For the present dataset, the tests decisively rejects the null hypothesis of unit roots for log(revenue), critic ratings, and user ratings. Hence, the panel data seems to be stationary and the fixed effects specification is preferred.

A random effects model assumes that the fixed effect (α_i in (M1)) is not correlated with each independent variable (Woolridge, 2013, p. 474). This is unlikely for the present dataset. On the contrary, concern about correlation between the fixed effect and ratings is one of the motivation for using fixed effects. One can use the Hausman test to explicitly check for correlation and to see which specification is better. For the following results, the Hausman test unequivocally rejects the null hypothesis that the unique errors are uncorrelated with the regressors. Hence, the decision for the fixed effects model seems well founded.

7 Results

This section presents the results. First, the models outlined in section 6 are tested with critic ratings as input (subsection 7.1.). Then, the same models are estimated using user ratings in subsection 7.2.

7.1 Critics

Table (3) presents the results from the fixed effects regressions using critic scores as the basis for ratings, as well as an OLS for comparison. The OLS estimate for critic ratings is negative, contrary to expectations. However, because the OLS likely suffers from omitted variables, a causal relationship cannot be concluded. Model 1, which seeks to establish whether critic ratings have an effect, has the coefficient for rating estimated at 0.0646 and is statistically significant at the 5% level. The point estimate would suggest that a 1-point increase (on the 0-100 scale) in average rating is associated with roughly 6.46% increase in revenue the next day, ceteris paribus.[9] The estimate is in this regard also economically significant. This is a rejection of the null hypothesis (H_1^0) and serves as evidence towards hypothesis (H_1). Number of reviews is also statistically and economically significant. An extra review is associated with a revenue increase of approximately 11.09%, which rejects the null and provides evidence for hypothesis (H_2). The estimate for the control variable of days since premiere has a significant negative sign and suggests that a movie loses on average about 5% of revenue each day.

[9] As the dependent variable is in log form, multiplying the coefficients by 100 approximates the percentage increase in a unit increase of the corresponding independent variable. For a more precise estimate, use $\%\Delta revenue = \left(e^{\beta} - 1\right) \cdot 100\%$

Table (3) – Fixed effects regressions

Dependent variable: log(revenue)

For critics

		OLS	Model 1	Model 2	Model 3	Model 4	Model 5
β_1	Rating	−0.0255***	0.0646**	0.0338	0.0085	0.0485	0.0880**
		(0.0005)	(0.0317)	(0.0283)	(0.0659)	(0.0341)	(0.0429)
β_2	Number of critic reviews	0.0963***	0.1105***	0.0928***	0.0933***	0.1108***	0.1108***
		(0.0006)	(0.0136)	(0.0128)	(0.0124)	(0.0135)	(0.0136)
β_3	Days since premiere	−0.0144***	−0.0508***	−0.1007***	−0.1007***	−0.0507***	−0.0508***
		(0.0005)	(0.0019)	(0.0061)	(0.0061)	(0.0019)	(0.0019)
β_4	Rating * days since premiere			0.0008***	0.0007***		
				(0.0001)	(0.0001)		
β_5	Rating * date				$1.79 \cdot 10^{-5}$		
					$(3.50 \cdot 10^{-5})$		
β_6	Ratings * Big Seven					0.1719**	
						(0.0722)	
β_7	Ratings * drama						−0.0697
							(0.0620)
	R−squared:						
	Within		0.6849	0.7202	0.7203	0.6853	0.6848
	Between		0.0305	0.0103	0.0047	0.2831	0.0467
	Overall	0.3508	0.1501	0.1246	0.0911	0.2526	0.1593
	Number of observations	127,147	126,412	126,412	126,412	126,412	126,283
	Number of movies	2,204	2,025	2,025	2,025	2,025	2,021[a]

Note: This table outlines the estimated coefficients for models 1-5 using ratings from critics. Model 1-5 control for movie and date fixed effects. Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively.

[a] 4 movies in the sample lack information on genre.

Having established that ratings do indeed have an effect, it remains to see how this effect may change in strength over various specifications. Model 2 evaluates whether a rating score has more to say later in a movie's lifecycle than at the premiere. The coefficient for rating * days since premiere (β_4) is positive and highly significant. Thus, ratings have more power for each day after a movie has been released. β_1 is in this specification not significant though. The point estimates suggest that on the premiere, an extra point in average rating increases the next day's revenue with circa 3.38%, whereas a week after, the effect is an increase in log(revenue) of $0.0338 + 7 \cdot 0.0008 = 0.0394$, or a roughly 3.94% increase in next day's revenue. The effect of ratings is thus about 17% stronger the following week according to the point estimates.

Model 3 estimates how ratings change over time. As the coefficient for rating * date tells (β_5), the relationship is not significant. Hence, the null hypothesis (H_5^0) is not rejected and evidence remains elusive for the claim that ratings strengthen over the period of study (2010-2018).

Model 4 shows a positive and significant estimate for the interaction between critic ratings and Big Seven. This mounts evidence to hypothesis H_5; movies produced by the Big Seven are thus more affected by a change in average ratings than other movies are. The point estimates suggest the effect is $0.1719/0.0485 = 3.54$ times as strong for Big Seven movies.

Model 5 is used to evaluate whether drama movies are more prone to ratings than other genres. However, as the estimate for rating * drama suggest, not only is there a lack of positive additional effect from belonging to the drama category, but the effect is negative although insignificant. Model 5 thus fails to reject hypothesis (H_7), and there is no evidence of drama movies in general being more prone to critic ratings, at least when considering critics in aggregate.

7.2 Users

Table (4) presents the fixed effects regressions for M1-M5 using user ratings. Looking at Model 1 with user ratings instead of critic ratings, the picture is dramatically different. The estimate for user ratings is negative, which would if interpreted literally suggest that log(revenue) goes down if average user ratings increase by 1 point. This is counterintuitive. An insignificant positive effect could be reconciled with the comparatively lower internet traffic on Metacritic, but a negative sign either suggests that consumers react contrary to the recommendations by Metacritic users or that the model has overlooked a relevant dynamic between user scores and revenue. This issue is tackled further in subsection 9.3.

Despite substituting critic ratings for user ratings, the qualitative interpretations of Model 2, Model 3, Model 4, and Model 5 is the same as in the case for critics. The effect of ratings increases as days since premiere passes. There is nothing to suggest that ratings generally strengthen over time. Big Seven movies are more sensitive to average rating changes than other movies are, and drama movies are not more influenced by ratings than other genres, at least not positively. The estimate for the interaction between drama and ratings for users is not only negative, but also statistically significant at 1% at that.

Table (4) – Fixed effects regressions

Dependent variable: log(revenue)

For users

		OLS	Model 1	Model 2	Model 3	Model 4	Model 5
β_1	Rating	0.0054***	−0.0068*	−0.0122**	0.0043	−0.0125**	0.0032
		(0.0004)	(0.0041)	(0.0049)	(0.0086)	(0.0053)	(0.0050)
β_2	Number of critic reviews	0.0024***	−0.0049***	−0.0050***	−0.0050***	−0.0048***	−0.0049***
		(0.0001)	(0.0015)	(0.0019)	(0.0019)	(0.0015)	(0.0015)
β_3	Days since premiere	−0.0155***	−0.0479***	−0.0787***	−0.0784***	−0.0474***	−0.0473***
		(0.0005)	(0.0020)	(0.0140)	(0.0140)	(0.0020)	(0.0020)
β_4	Rating * days since premiere			0.0005**	0.0005**		
				(0.0002)	(0.0002)		
β_5	Rating * date				$-1.14\cdot10^{-5}$**		
					$(4.68\cdot10^{-6})$		
β_6	Ratings * Big Seven					0.0200**	
						(0.0079)	
β_7	Ratings * drama						−0.0220***
							(0.0085)
	R−squared:						
	Within		0.6998	0.7133	0.7139	0.7003	0.7002
	Between		0.0137	0.0188	0.0180	0.0001	0.0065
	Overall	0.2094	0.0734	0.0723	0.0794	0.1118	0.0853
	Number of observations	121,429	120,710	120,710	120,710	120,710	120,598
	Number of movies	1,834	1,754	1,754	1,754	1,754	1,752[a]

Note: This table outlines the estimated coefficients for models 1-5 using ratings from users. Model 1-5 control for movie and date fixed effects. Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively.

[a] 2 movies in the sample lack information on genre.

8 Robustness Checks

This section re-estimates the models outlined in the previous section with certain modification and assesses the sensitivity with respect to additional controls.

8.1 First differencing

Even though subsection (6.5) argues for the use of fixed effects as the main specification, a comparison to the first difference regression can still be made. The full results are reported in the appendix, with Table (8) using critics as basis for rating and Table (9) using users as basis for rating. For critics, all coefficients are rendered insignificant with the exception of number of reviews, which remain highly significant and quantitatively in line with the fixed effects estimates at β_2= 0.1176. The results are overall therefore sensitive to the choice of model. The lower estimated effect in the first difference regression (β_1=0.0101) is probably because the first difference regression estimates the day to day effect and commands the effect be immediate. Time-demeaning the data through the within transformation on the other hand, manages to capture delays in the influencing effect when variation in ratings is zero for several days in a row. This can explain why the estimates differ.

For users, the estimate for ratings is positive and highly significantly so. The point estimate for β_1 = 0.0008, suggests that an increase in average user rating raises the next days revenue by roughly 0.08%. The estimate for the interaction between days since premiere and ratings is also highly significant. The estimate for the date interaction remains insignificant. For the Big Seven, the model estimates a negative additional effect, contrary to the results obtained by the fixed effects estimate.

8.2 Exclusion criteria

Observations that occurred more than 1 year after the premiere are excluded from the analysis to avoid these outliers to skew the coefficients. To see how sensitive the results are to this change, Models 1-5 can be re-estimated including all observations (full set of results are reported in Table (10) for critics and Table (11) for users, both found in the appendix). The within R-squares are lower across the board in this specification than the original. The re-estimated models thus fits the data worse, and the exclusion criteria seems well founded. Thus, the original specifications ought to have a higher bearing for resting conclusions on. Still, the qualitative conclusions of all hypotheses are unaffected, with the exception of number of reviews, which is rendered insignificant.

8.3 More movies over time

The lack of significant effect for the rating * date variable could be due to more number of movies also influencing rating. The number of movies has generally increased over the years. If ratings behave differently depending on the number of movies that are available on the market, then this

can bias the estimate for β_5 in Model 3. To test for this possibility, Model 3 is modified and re-estimated. The full results are under the column for Model 7 in Table (5) in the appendix, but the short version is that the inclusion of this control does not turn the estimate for β_5 significant. As a side note there seems to be no relationship in how ratings change in effect depending on the number of movies running on a given day.

8.4 Slower strengthening of ratings over time

Another possible explanation for failing to detect a significant effect in the β_5 estimate for rating * date is that Model 3 too aggressively amplifies ratings over time. An alternative is to let the time variable be of lower frequency. For instance, instead of interacting rating with a date variable that increases each day, one can interact rating with a year variable, which only increases each year. This enables ratings to strengthen more slowly over time. As there are exactly eight years in the dataset, a year variable ranging from 1-8 is constructed, starting on 22 October 2010. Table (7) in the appendix shows the estimates of this alternative specification for both critics and users. For critics, the estimate remains insignificant, albeit changing sign from positive to negative. For users, the estimate remains negative and statistically significant. This alteration does therefore not change any of the conclusions.

8.5 Drama movies correlated with Big Seven

Although drama movies in general are no more inclined to changing ratings than other genres, a possible explanation for this might stem from drama movies correlating negatively with Big Seven movies. Re-estimating Model 4 with the inclusion of an interaction between critic ratings and Big Seven (results under column for Model 6 in Table (5) in the appendix) does however not change the conclusion. The estimate is still negative at −0.0620, and significant at the 1% level.

8.6 Simultaneity in number of critic reviews and revenue

The thesis did not postulate a hypothesis of the awareness effect for users, since number of reviews is likely endogenous with revenue. For critics, one may also raise legitimate objections on grounds of endogeneity. One possibility is that of simultaneity. If critics post reviews in response to anticipated traffic at the cinemas, then the estimate of β_2 in Model 1 is biased. When having two simultaneous equations with the dependent variable of one being the independent of the other and vice versa, a tool to disentangle the causal effects is the use of instrumental variables. To find out the causal effect of number of reviews on log(revenue), an instrument for number of reviews must be identified. This instrument must correlate with number of reviews but not with log(revenue). In the present dataset, no such viable instrument exist. However, as a second best, one might estimate what the causal effect of log(revenue) is on the lagged version of number of reviews. This might seem counterfactual, as a variable cannot have a casual effect on another variable back in time. The appropriate interpretation is however that of expected revenue. If number of reviews is a function of expected revenue, then a "causal" effect interpretation exists. Having established critic ratings to affect revenue, it ticks the relevance criterion for instruments. The time variation of ratings is

also uncorrelated with the time variation for number of reviews. Hence, instrument exogeneity is also present and critic ratings may be used as an instrument. This is essentially a check to see whether number of reviews is a function of expected future revenue. The full results of this is reported in Table (12) in the appendix, but the conclusion remains that there is no evidence that number of reviews predicts sales, with the p-value of the coefficient being 73%.

9 Discussion

This section discusses the results obtained. First comes an evaluation on the reliability of the causal effect interpretation within the confines of the given dataset (subsection 9.1). Subsection 9.2 then provides some arguments for why the results are likely to hold for other contexts. Subsection 9.3 is dedicated towards the unanticipated negative estimate of user ratings in influencing revenue and explores possible resolutions. Subsection 9.4 looks at the results in conjunction with previous literature. The utility of the results for the manager and the policy maker is addressed in subsection 9.5 and 9.6 respectively.

9.1 Internal validity

The design of the econometric strategy removes a lot of concern for potential biases in the result. All the unique attributes that a movie possesses and which does not change over time is automatically controlled for. This reduces much of the worry down to potential variables and mechanisms that simultaneously correlate with ratings and revenue changes. Although advertising is such a candidate in biasing the estimated coefficients for the rating variables, it likely has little impact. As explained in the Econometric Strategy section, the industry practices in the movie industry severely limits the flexibility in adjusting advertising through a movie's lifecycle. Furthermore, the bulk of advertising is spent ahead of release, which only leaves the remaining minority to be spent in the post-release period that is the basis for this study. Having advertising expenditure as a control variable would certainly not hurt, but the exclusion of it should not undermine the results, which estimates ratings to have a large positive effect.

For users, there might be more concerns for endogeneity. Since the identity of users are not known and they are not paid to provide a critique of any movie, the motivation of some users could interfere with how revenue is affected over time. Subsection 9.3 discusses this further.

Apart from ratings, the estimated awareness effect might have endogeneity issues. For users, this seems very likely and the effect is consequently not tested for. Number of user reviews may very well be a proxy for how many people have seen it thus far, and by extension, a proxy of a depleting potential customer base. It is a critic's job to review movies on the other hand, and so the same mechanism that underlies when a user posts a review is likely different from when a critic does so. Still, if the time at which a critic posts a review is somehow a function of an unobserved characteristic affecting revenue, the coefficient might be biased. For instance, if critics systematically publish their reviews before a day where it is expected many visitors to the cinemas,

then the estimated effect could be partly a measure of the predicting power reviews have in anticipating sales. However, as subsection 8.5 outlines, no such evidence exist.

It is also possible that new reviews are produced in response to higher revenue in the past. As far as revenue being higher for some movies due to a set of movie characteristics and attracting more reviews, this should be part of the fixed effect and thus out of concern. However, if revenue is increasing over time for a given movie, and a new review comes out in the middle of this upward trend, it can interfere with the interpretation of the estimate. Nonetheless, this would have the same effect as a bias for future expected revenue, and as the previous paragraph has highlighted, there is no evidence of simultaneity between number of reviews for a given movie, and revenue.

9.2 External validity

Of interest is the extent to which the results obtained in this study can be generalised to other markets. This can both be other film markets (i.e. outside USA and Canada) as well as other products than movies. For both of these classifications, the conclusions that critic ratings positively affect revenue and that more reviews raise sales, is likely to hold for many other products. The overarching argument for this is that there is nothing specific about the US and Canada or movies as a product that the underlying theoretical models or introspection hinges upon to make the hypotheses. This is not to say that the effect is likely to be the same for other markets. It most certainly is not. Nonetheless, the qualitative implications are hypothesised to exist independent of culture and product. The sample is thus probably representative in direction but not in magnitude for a host of products in general. Despite this, there are probably instances where one might expect the result not to hold. The cash flow for monopolies of a highly inelastic good might be close to unaffected by any ratings or number of reviews. Still, instances where ratings are not important for sales are probably exceptions.

Ratings strengthening over a product's lifetime may however not hold for markets generally. Films are for many a one-shot purchase. Learning from self is thus less valuable than for a repeat-purchase good; regardless of whether a customer likes a movie, he will not continue purchasing new tickets anyway making the information that he liked or disliked the movie of limited value. By substitution, learning from others may therefore have more value in such an environment. Hence, delaying purchase to learn from others can have a wider spread in a film market than in markets for repeat-purchase goods. Another difference is that markets for many products and services do not have the same constraints on time as that of cinemas. The fact that films are introduced, stays on the screens for some weeks and then taken off may make the strengthening effect of ratings as days since premiere passes more salient. Even if such an effect is general for consumer products, it is likely that it withers off over time. Nevertheless, the strengthening effect over a lifecycle is still likely to hold for other film markets, and may hold for other markets of goods that by many are purchased once and only once, such as video games and books of fiction.

9.3 Why are positive user ratings estimated to have a negative effect?

It is puzzling not only that favourable user ratings lack a positive effect, but also that Model 1 estimates a negative influence of the value of user ratings on sales. It is possible that the model has overlooked an important dynamic in how the ratings of user reviews are determined. One possibility is that extreme negative movements away from the consensus reflect controversies, fuelling word-of-mouth, which increase the awareness of a film and in turn attract an audience. Another possibility might lie in users reacting differently to certain movies that have received praise from critics.

Judging by the reviews of some movies, there is often a discrepancy between the average critic and user ratings. Although the correlation between the time-demeaned versions of critic and user ratings is very low (0.01), there might still be a dynamic in how users are deciding their rating based on the overall critic scores. To test this conjecture, Model 1 for users is re-estimated with an interaction term between critic and user ratings, while controlling for critic ratings. Employing this specification reveals an estimate for average user rating at 0.0529 (se = 0.0128, two-tailed p-value < 0.0001) on log(revenue), whereas the interaction term is negative at −0.0009 (se = 0.0002, two-tailed p-value < 0.0001). Full results are included in Table (6) in the Appendix. Although the estimate for user rating is positive and highly significant, the interaction between critic and user ratings is negative. Thus, for movies with a sufficiently low critic rating, positive user ratings do seem to influence sales favourably, but not so for movies with a high critic rating. The point estimates suggest that user ratings positively influence revenue when the average critic rating is below 59.[10] This is close to the average among all movies, which is 61 (see Data section). One possibility is that user and critic ratings serve different types of consumers and signal different aspects of a movie. Insofar as a higher user rating signals aspects of a movie that one consumer segment likes, but not another, it is not impossible that user ratings may have a negative effect on sales for some movies, if the consumer segment that does not infer good quality from high user ratings dominates.

An alternative explanation is that users might try to retaliate with bad ratings for movies they consider overrated by critics. As a case study to this is the most financially successful movie in the dataset, namely "Star Wars: The Force Awakens", which has an average critic score of 80.91 but an average user score of 55.92. The user score decreased from its starting point of 67.17 on the premiere. The ensuing down voting may however not produce a strong enough effect to compensate for the favourable critic ratings. It thus produces an illusion that user reviews do not matter, whereas in reality they might. This is one explanation, but further study on the relationship between users and critics is welcome.

[10] The estimated effect on log(revenue) from a change in user rating is $0.0529 - 0.0009 * \text{critic}$, where critic refers to average critic rating. Setting this as an inequality to zero and solving for critic reveals that the effect is positive when $\text{critic} < 0.0529 / 0.0009 \approx 59$.

9.4 Comparison with previous literature

Critics having a critical role in influencing economic decisions is in accordance with the study of Friberg and Grönqvist (2012), whose study focused on the demand for wine, and Berger et al. (2010), whose study revolved around books. The more direct comparison, however, might be made to Reinstein and Snyder (2005), who also study critics' effect on box office receipts. Unlike the results in this study, Reinstein and Snyder (2005) only find a marginal significant effect of critic ratings on revenue. Differences in methods and data may explain why the conclusions do not match. This study has higher frequency of the data, with daily observations, whereas Reinstein and Snyder (2005) used weekend data with two observations per movie. Furthermore, Reinstein and Snyder (2005) study the effect of two critics. This study looks at 90 media houses. Finally, the sheer magnitude of this dataset is likely to have a higher probability of detecting an influencing effect. Hence, the lack of strong significance in Reinstein and Snyder (2005) may be a false-negative. It is worth mentioning that the result is significant at the 10% level, so coupled with this study it seems probable that critic ratings do indeed influence sales.

User reviews for the film industry are as the results of Duan et al. (2008) not found to have an effect. As the previous subsection (9.3) demonstrated however, the relationship might be more complicated. Hence, overconfident assessments of a lacking effect from anonymous users warrant caution and further research may help to paint a clearer picture. Looking beyond the film industry however, the effect of users (or consumers) has shown to be influential. In particular, the popularity of dishes by consumers in a Chinese restaurant chain had an influential effect for other consumers as demonstrated in Cai et al. (2009) for instance. This was however in another market and the nature of the rating differed. Instead of representing a sample saying how much they liked a certain dish, the "ratings" were rather that of indicating popularity.

The evidence for hypothesis (H_4) also provides an alternative explanation to the results of the pioneering paper by Eliashberg and Shugan (1997). Correlation being more significant in the later weekends of a movie being screened but not in the opening weekends does not imply critics serving more as predictors than influencers. Instead, ratings may simply increase in importance over the lifecycle. The explanation could be in accordance with the motivation for the hypothesis. That is, certain customers who have weaker priors that they will like a movie delay the ticket purchase until more learning from others can be gleaned.

The results of Berger et al. (2010) indicate that an unfavourable book review is bad for a famous author but still beneficial for an unknown author. Thus, awareness might be more important than ratings for unknown products. This study finds more reviews to boost sales, and this might be a result of the awareness effect. The results also point to a significant additional influencing effect from ratings by critics for movies distributed by the Big Seven, serving as proxy for more famous movies. In tandem therefore, these results suggest that the actual value of ratings matter more for sales the better known a product is.

The study fails to replicate the findings of Reinstein and Snyder (2005). The 90 critics considered do not have more power on drama movies than on other genres. This does not necessarily mean that Reinstein and Snyder's (2005) findings are entirely incorrect, as the study is on two critics in

particular. Some critics might have more influence on certain arenas than others. Still, the results cast a shadow of doubt that the findings are robust. The point estimate is even negative for the results obtained in the present thesis. One thing that seems safe to conclude at least, is that there is no strong evidence of critics in general having additional influence on the revenue stream of drama movies compared to other genres.

9.5 Managerial implications

Although in the dataset considered, the correlation between ratings and revenue is very low and thus explains little of the variation (see Data section), further examination shows that for any given movie, ratings do have a sizeable effect on revenue. Therefore, a profit-maximising business should not ignore them, but rather aim to increase them if the cost is sufficiently low. Hiring critics in the production process to rate a product pre-release or even to rate a set of concepts or ideas pre-production is a concrete advice.

9.6 Policy implications

As ratings guide economic behaviour, manipulation thereof may distort welfare. A policy maker may thus consider laws that prevent unscrupulous manipulation of ratings that does not arise from improving the underlying product. Although judicial obstacles may prevent the extent to which such laws can be implemented, certain policies could be more realistic than others. For instance, if it can be proven beyond reasonable doubt that a company has conspired to manipulate ratings by say, bribes to critics, then this could be decreed punishable. It is possible that detection of such behaviour would be sanctioned naturally in a laissez-faire environment, but such behaviour could still compromise the wider integrity of the ratings platform, thus serving as a negative externality.

10 Limitations and Further Research

This thesis pools critic ratings into an average measure and estimates the effect. The effect can both be due to people looking at the averages through influential sites such as Metacritic, IMDb, and Google, but it can also be due to the constituent ratings themself. It is not possible to disentangle these to see how much of the influence is due to the average and which is due to independent reviews. The scope of this thesis is to see whether ratings have an effect, not how this effect is dispersed through the various channels. A future study might however look more on how the effect manifests itself. It is also possible that it is not the numerical ratings per se influencing the consumer, but rather the underlying review (i.e. the text).

Ratings for a given movie increase the demand for cinema-tickets for that movie, but this study does not look at the ripple effects that ratings can have beyond the confines of box office revenue. It seems likely that if ratings affect box office revenues, they also affect the sales for home entertainment (e.g. DVD, Blu-ray) versions of those movies. Since the effect of ratings seems to

strengthen over the product's life, the effect of rating may be even greater for home entertainment. Furthermore, good ratings may also have a spill over effect on prequels or sequels or other movies produced by the studio or director. A future similar study may analyse possible spill over effects.

This paper uses Metacritic users to study the effect of ratings by consumers of the product, but as previously mentioned this is not the biggest forum for evaluating movies; IMDb is. A future study could look at the effects that users from different forums have in influencing revenue. This could also preferably be with data on all ratings, not just ratings associated with written reviews. This study has also implicitly assumed that all ratings are created equal. This assumption is likely false, yet does not undermine the qualitative findings. Still, there is value in knowing the differences in influence by reviewers. For critics, one approach could be a scaled up version of Reinstein and Snyder's (2005) methods.

The negative estimate of user ratings on log(revenue) opens up avenues for further research. From subsection 9.3 it seems that an important dynamic between users and critics are overlooked by the main analysis. After controlling for user and critic reviews, there is a negative interaction between them. What drives this effect is an interesting problem that a researcher can attempt to solve.

Both this study and Berger et al. (2010) underscore the importance of the awareness effect. Since there is little research on this area, exploration of theories awaits. It is for instance possible to modify the theoretical model of e.g. McFadden and Train (1996) to include awareness as an input. It seems likely, that a small entrant can take bigger risks in not appealing to the mass audience if this serves to increase attention. Hence, this might serve as a remedy to the convergence prediction produced by the model (ibid.). It is also possible that awareness might come through controversial PR stunts. It is better to have 100,000 people aware of a product with a probability of 1% that each person engages in a purchase, than to have 100 people whose same probability is 90%. One might also wonder how general this effect is beyond the realm of economics. For instance, maybe it extends to politics. Some established political parties might have more to lose from controversies than a newcomer might. If a fresh political party enters the scene, it may benefit from preaching extreme views even if those views cater to maximum 5% of the population, should this create a lot of attention and media coverage in the process.

11 Conclusion

This thesis expands a small literature on the relationship between ratings and sales. Previous studies mostly find that positive reviews influence sales favourably, but data sets from the film industry have suggested that there is no significant effects (Reinstein and Snyder, 2005; Duan et al., 2008). With the results found in this study, which employs a much larger data set, there should be less doubt that higher ratings from critics do in fact contribute to higher sales. Interestingly, user ratings are estimated to work inversely with sales revenue. Some exploratory analysis reveals that when controlling for critic scores, user scores do have positive effects as well for movies with low critic scores, but hard conclusions on the effect of users should be delayed for a future study. The study also find more reviews to increase sales, which might be due to raising awareness.

The novelty in this thesis lies in the studying of how ratings may change in its effect. Ratings matter more the longer the movie has run in the cinemas, but one cannot conclude that there has been a general trend in the film industry of ratings strengthening in importance over the years. Thus, the proliferation of new IT technologies such as smartphones and higher internet penetration, enabling consumers to consult others for information about which products to buy, or in this case, which movies to watch, has not shown to significantly shift consumer behaviour on this front.

Movies from the seven major film studios that control 80-90% of the US and Canadian film market are more affected by average ratings than their smaller competitors are. This is probably partly due to more awareness of these movies. Finally, drama movies do not seem to be more affected by critic ratings than other genres; if anything, they have less to say.

References

Alexa (2018a). *Imdb.com Traffic Statistics*. URL: https://www.alexa.com/siteinfo/imdb.com [last accessed 29 November 2018].

Alexa (2018b). *Metacritic.com Traffic Statistics*. URL: https://www.alexa.com/siteinfo/metacritic.com [last accessed 26 November 2018].

Banerjee, A.V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics* 107.3, pp. 797-817.

Berger, J., Sorensen, A.T., and Rasmussen, S.J. (2010). Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science* 29.5, pp. 815-827.

Cai, H., Chen, Y., and Fang, H. (2009). Observational Learning: Evidence from a Randomized Natural Field Experiment. *American Economic Review* 99.3, pp. 864-882.

Duan, W., Gu, B., and Whinston, A.B. (2008). Do Online Reviews Matter? – An Empirical Investigation of Panel Data. *Decision Support Systems* 45, pp. 1007-1016.

Elberse, A. and Anand, B. (2007). The Effectiveness of Pre-release Advertising for Motion Pictures: an Empirical Investigation Using a Simulated Market. *Information Economics and Policy* 19, pp. 319-343.

Eliashberg, J. and Shugan , S.M. (1997). Film Critics: Influencers or Predictors? *Journal of Marketing* 61.2, pp. 68-78.

Friberg, R. and Grönqvist, E. (2012). Do Expert Reviews Affect the Demand for Wine? *American Economic Journal: Applied Economics* 4.1, pp. 193-211.

Granger, C.W.J. and Newbold, P. (1974). Spurious Regressions in Econometrics. *Journal of Econometrics* 2, pp. 111-120.

Greene, W. (2000). *Econometric Analysis*. 4th edition. Upper Saddle River, New Jersey: Pearson Prentice-Hall.

Hendricks, K. and Sorensen, A. (2009). Information and the Skewness of Music Sales. *Journal of Political Economy* 117.2, pp. 324-369.

Maddala, G.S. and Wu, S. (1999). A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. *Oxford Bulletin of Economics and Statistics* 61 (Special Issue Nov), pp. 631-652.

McFadden, D.L. and Train, K.E. (1996). Consumers' Evaluation of New Products: Learning from Self and Others. *Journal of Political Economy* 104.4, pp. 683-703.

Motion Picture Association of America (2018). *THEME Report: A Comprehensive Analysis and Survey of the Theatrical and Home Entertainment Market Environment (THEME) for 2017*. URL: https://www.mpaa.org/wp-content/uploads/2018/04/MPAA-THEME-Report-2017_Final.pdf [last accessed: 26 November 2018].

Orbach, B.Y. and Einav, L. (2007). Uniform Prices for Differentiated Goods: The Case of the Movie-Theater Industry. *International Review of Law and Economics* 27.2, pp. 129-153.

Reinstein, D.A. and Snyder, C.M. (2005). The Influence of Expert Reviews on Consumer Demand for Experience Goods: a Case Study of Movie Critics. *The Journal of Industrial Economics* 53.1, pp. 27-51.

Schatz, T. (1999). *Boom and Bust: American Cinema in the 1940s*. Oakland, California: University of California Press.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22.11, pp. 1359-1366.

Sissors, J.Z. and Baron, R.B. (2010). *Advertising Media Planning*. 7th ed. New York City, New York: The McGraw-Hill Companies, Inc.

Smallwood, D.E. and Conlisk, J. (1979). Product Quality in Markets Where Consumers are Imperfectly Informed. *The Quarterly Journal of Economics*, 93.1, pp.1-23.

Vogel, H.L. (2007). *Entertainment Industry Economics: A Guide for Financial Analysis*. 7th ed. New York City, New York: Cambridge University Press.

Woolridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. London: The MIT Press.

Woolridge, J.M. (2013). *Introductory Econometrics: A Modern Approach*. 5th ed. New Dehli: South-Western, a part of Cengage Learning.

Appendix

A.1 List of media houses providing critic reviews used in this study

Arizona Republic
Austin Chronicle
Baltimore Sun
Boston Globe
Boxoffice Magazine
Charlotte Observer
Chicago Reader
Chicago Sun-Times
Chicago Tribune
Christian Science Monitor

CineVue
Consequence of Sound
Dallas Observer
Empire
Entertainment Weekly
Film Journal International
Film Threat
Film.com
Hitfix
IGN

Indiewire
L.A. Weekly
LarsenOnFilm
Los Angeles Times
MTV News
McClatchy-Tribune News Service
Miami Herald
Movie Nation
MovieLine
Mr. Showbiz

NPR
New Orleans Times-Picayune
New Times (L.A.)
New York Faily News
New York Magazine (Vulture)
New York Post
Newsweek
Observer
Original-Cin
Orlando Sentinel

Paste Magazine
Philadelphia Daily News
Philadelphia Inquirer
Portland Oregonian
Premiere
ReelViews
RogerEbert.com
Rolling Stone
Salon
San Francisco Chronicle

San Francisco Examiner
Screen International
ScreenCrush
Seattle Post-Intelligencer
Slant Magazine
Slate
St. Louis Post-Dispatch
TNT RoughCut
TV Guide Magazine
Tampa Bay Times

The A.V. Club
The Associated Press
The Atlantic
The Dissolve
The Film Stage
The Globe and Mail (Toronto)
The Guardian
The Hollywood Reporter
The New Republic
The New York Times

The New Yorker
The Observer (UK)
The Playlist
The Seattle Times
The Telegraph
The Verge
TheWrap
Time
Time Out
Time Out London

Total Film
USA Today
Uproxx
Vanity Fair
Variety
Village Voice
Vox
Wall Street Journal
Washington Post
We Got This Covered

A.2 Additional tables

Table (5) – Fixed effects regressions with additional controls

For critics

		Dependent variable: log(revenue)	
		Model 6	**Model 7**
β_1	Rating	0.0700	0.0194
		(0.0454)	(0.0698)
β_2	Number of reviews	0.1111***	0.0820***
		(0.0507)	(0.0017)
β_3	Days since premiere	−0.0507***	−0.0976***
		(0.0019)	(0.0000)
β_4	Rating * days since premiere		0.0007***
			(0.0000)
β_5	Rating * date		$-1.16\cdot10^{-5}$
			$(3.61\cdot10^{-5})$
β_6	Rating * Big Seven	0.1644***	
		(0.0742)	
β_7	Rating * drama	−0.0620	
		(0.0612)	
β_8	Number of movies		0.0080
			(0.0111)
β_9	Rating * number of movies		−0.0002
			(0.0002)
	R−squared:		
	Within	0.6852	0.6729
	Between	0.2774	0.0353
	Overall	0.2471	0.2265
	Observations	126,283	126,412
	Number of movies	2,021[a]	2,025

Note: The table outlines the fixed effects regressions used as robustness checks from section 8. Standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively

[a] 4 movies in the sample lack information on genre.

Table (6) – Fixed effects regression with interaction between critic and user

	Dependent variable: log(revenue)
	Model 8
Rating (critic)	0.1529***
	(0.0405)
Rating (user)	0.0529***
	(0.0128)
Rating (critic * user)	−0.0009***
	(0.0002)
Number of critic reviews	0.1143***
	(0.0140)
Number of user reviews	− 0.0049
	(0.0016)
Days since premiere	−0.0518
	(0.0023)
R−squared:	
Within	0.7288
Between	0.0029
Overall	0.0993
Observations	116,625
Number of movies	1,652

Note: The table estimates the model of subsection 9.3.
Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively.

Table (7) – Fixed effects regression with interaction between rating and year

	Dependent variable: log(revenue)	
	Critics	**Users**
Rating	0.0523	0.0048
	(0.0375)	(0.0064)
Number of reviews	0.0922***	−0.0050***
	(0.0129)	(0.0019)
Days since premiere	−0.1007***	−0.0788***
	(0.0061)	(0.0140)
Rating * days since premiere	0.0008***	0.0005**
	(0.0001)	(0.0002)
Rating * year	−0.0052	−0.0044***
	(0.0053)	(0.0012)
R-squared:		
Within	0.7203	0.7142
Between	0.0160	0.0175
Overall	0.1497	0.0785
Number of observations	126,412	120,709
Number of movies	2,025	1,753

Note: The table estimates the model as referred to in subsection 8.5. The year variable ranges from 1-8, starting with 1 for the year starting on 22 October 2010.
Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively.

Table (8) – First difference regressions

For critics		Dependent variable: log(revenue)				
		FD 1	**FD 2**	**FD 3**	**FD 4**	**FD 5**
β_1	Rating	0.0101 (0.0086)	0.0005 (0.0088)	−0.0103 (0.0277)	0.0111 (0.0092)	0.0093 (0.0109)
β_2	Number of critic reviews	0.1176*** (0.0108)	0.1174*** (0.0108)	0.1178*** (0.0105)	0.1174*** (0.0108)	0.1175*** (0.0108)
β_4	Rating * days since premiere		0.0004 (0.0004)	0.0004 (0.0004)		
β_5	Rating * date			$9.98{\cdot}10^{-6}$ $(1.48{\cdot}10^{-5})$		
β_6	Ratings * Big Seven				−0.0093 (0.0259)	
β_7	Ratings * Drama					0.0026 (0.0180)
	R−squared:	0.6038	0.6032	0.6032	0.6032	0.6032
	Number of observations	122,961	122,961	122,961	122,961	120,872

Note: This table outlines the estimated coefficients for the first difference version of Models 1-5 as specified in section 6 using critics as basis for ratings. The models control for date fixed effects through dummies (not reported here).
Robust standard errors in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively

Table (9) – First difference regressions

For user		Dependent variable: log(revenue)				
		FD 1	**FD 2**	**FD 3**	**FD 4**	**FD 5**
β_1	Rating	0.0008***	0.0007***	0.0006***	0.0009***	0.0007***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
β_2	Number of critic reviews	−0.0010***	−0.0008***	−0.0008	−0.0009***	−0.0009***
		(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
β_4	Rating * days since premiere		$2.55 \cdot 10^{-6}$***	$2.56 \cdot 10^{-6}$***		
			$(4.58 \cdot 10^{-7})$	$(4.58 \cdot 10^{-7})$		
β_5	Rating * date			$7.32 \cdot 10^{-8}$		
				$(8.58 \cdot 10^{-8})$		
β_6	Ratings * Big Seven				$-1.06 \cdot 10^{-4}$***	
					$(3.13 \cdot 10^{-5})$	
β_7	Ratings * Drama					0.0002***
						$(3.68 \cdot 10^{-5})$
	R−squared:	0.6640	0.6642	0.6642	0.6641	0.6642
	Number of observations	114,135	114,135	114,135	114,135	114,025

Note: This table outlines the estimated coefficients for the first difference version of Models 1-5 as specified in section 6 using users as basis for rating. The models control for date fixed effects through dummies (not reported here).
Robust standard errors in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively

Table (10) – Fixed effects regressions without exclusion criteria

For critics		Dependent variable: log(revenue)				
		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
β_1	Rating	0.0925***	0.0399	0.0302	0.0754**	0.1249***
		(0.0360)	(0.0262)	(0.0555)	(0.0380)	(0.0474)
β_2	Number of critic reviews	0.0313	0.0221	0.0223	0.0317	0.0318
		(0.0282)	(0.0260)	(0.0261)	(0.0281)	(0.0281)
β_3	Days since premiere	−0.0358***	−0.1065***	−0.1065***	−0.0358***	−0.0358***
		(0.0050)	(0.0079)	(0.0079)	(0.0050)	(0.0050)
β_4	Rating * days since premiere		0.0010***	0.0010***		
			(0.0001)	(0.0001)		
β_5	Rating * date			$6.84 \cdot 10^{-6}$		
				$(3.01 \cdot 10^{-5})$		
β_6	Ratings * Big Seven				0.1830**	
					(0.0770)	
β_7	Ratings * Drama					−0.0969
						(0.0607)
	R−squared:					
	Within	0.5250	0.6069	0.6069	0.5255	0.5251
	Between	0.0020	0.0051	0.0056	0.1373	0.0002
	Overall	0.0203	0.0223	0.0199	0.1343	0.0310
	Number of observations	127,147	127,147	127,147	127,147	127,018
	Number of movies	2,025	2,025	2,025	2,025	2,021[a]

Note: This table outlines the estimated coefficients for models 1-5 specified in section 6 using ratings from critics without imposing the exclusion criterion for observations after one year following the premiere of a movie. Models 1-5 all control for movie and date fixed effects.
Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively.
[a] 4 movies in the sample lack information on genre.

Table (11) – Fixed effects regressions without exclusion criteria

		Dependent variable: log(revenue)				
For user		**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
β_1	Rating	0.0047 (0.0080)	−0.0082 (0.0105)	−0.0010 (0.0190)	0.0045 (0.0082)	0.0150 (0.0101)
β_2	Number of critic reviews	−0.0148** (0.0060)	−0.0123** (0.0054)	−0.0124** (0.0002)	−0.0149** (0.0060)	−0.0149** (0.0060)
β_3	Days since premiere	−0.0248*** (0.0072)	−0.0696** (0.0279)	−0.0695** (0.0280)	−0.0248*** (0.0072)	−0.0248*** (0.0072)
β_4	Rating * days since premiere		0.0006* (0.0003)	0.0006* (0.0003)		
β_5	Rating * date			$5.02 \cdot 10^{-6}$ $(7.63 \cdot 10^{-6})$		
β_6	Ratings * Big Seven				0.0007 (0.0120)	
β_7	Ratings * Drama					−0.0228** (0.0091)
	R−squared:					
	Within	0.4591	0.5289	0.5290	0.4591	0.4597
	Between	0.0705	0.00766	0.0709	0.0673	0.0376
	Overall	0.0231	0.0470	0.0465	0.0242	0.0343
	Number of observations	121,429	121,429	121,429	121,429	121,317
	Number of movies	1,754	1,754	1,754	1,754	1,752[a]

Note: This table outlines the estimated coefficients for models 1-5 specified in section 6 using ratings from users without imposing the exclusion criterion for observations after one year following the premiere of a movie. Models 1-5 all control for movie and date fixed effects.
Robust standard errors clustered by movies in parentheses. ***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively

[a] 2 movies in the sample lack information on genre.

A.3 Robustness check of simultaneity between number of reviews and revenue

Consider the simultaneous equations system:

(A1) $$\log(revenue_{i,t}) = \gamma_{11} rating_{i,t-1} + \gamma_{12} Nreviews_{i,t-1} + \gamma_{13} sinceprem_{i,t} + \alpha_i + u_{i,t}.$$

(A2) $$Nreviews_{i,t-1} = \gamma_{21} E_{t-1}[\log(revenue_{i,t})] + \gamma_{22} sinceprem_{i,t} + \alpha_i + u_{i,t}.$$

Even though expected revenue deviations cannot be observed, the actuals can be used as proxy. $rating_{i,t-1}$ is used as an instrument for $\log(revenue_{i,t})$ in (A2). The results are reported in Table (12). In Panel (2), one can see the estimate for $\log(revenue_{i,t})$ is negative yet insignificant from zero. Thus, there is no evidence that critics post their reviews in expectation to higher revenue.

Table (12) – Instrumental variables fixed effects regression to test for simultaneity

Panel (1) – First stage within regression:	
	Dependent variable: $\log(revenue_{i,t})$
$rating_{i,t-1}$	0.0791**
	(0.0331)
$sinceprem_{i,t}$	−0.0448***
	(0.0023)
R-squared:	
Within	0.5751
Between	0.0464
Overall	0.1132
Panel (2) – Fixed effects (within) IV regression	
	Dependent variable: $Nreviews_{i,t-1}$
$\log(revenue_{i,t})$	−1.228
	(3.2211)
$sinceprem_{i,t}$	0.2638*
	(0.1420)
R-squared:	
Within	0.1717
Between	0.0135
Overall	0.0102

Note: Robust standard errors clustered by movies in parentheses. $Nreviews_{i,t-1}$ refers to number of reviews from which ratings are drawn, and $sinceprem_{i,t}$ refers to days since premiere.
***, **, and * signify statistical significance at the 1%, 5%, and 10% respectively