Aguiar, Luis

Working Paper

Bad Apples on Rotten Tomatoes: Critics, Crowds, and Gender Bias in Product Ratings

CESifo Working Paper, No. 11422

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

CESifo WORKING PAPERS

11422
2024

October 2024

Bad Apples on Rotten Tomatoes: Critics, Crowds, and Gender Bias in Product Ratings

Luis Aguiar

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11422*

Bad Apples on Rotten Tomatoes: Critics, Crowds, and Gender Bias in Product Ratings

Abstract

Consumers considering the purchase of experience goods can rely on both critics and crowd-based evaluations to guide their decisions. Due to divergent incentives, however, critics and crowd assessments may incorporate different information. In the context of the movie industry, I investigate whether crowd reviewers provide gender-neutral product evaluations relative to professional critics. I classify movies as male or female based on the gender composition of their cast and estimate how the gender gap in movie rating scores differs across critics and crowds. Results show that while critics tend to assess both male and female movies similarly, the gender gap in ratings increases substantially under crowd-based ratings and at the expense of female movies. Notably, female movies receive a higher proportion of extreme low ratings, predominantly from male crowd reviewers. Using a rating design change implemented by the review-aggregating platform Rotten Tomatoes, results indicate that this overall increase in gender inequality is driven by a selected group of online reviewers rather than by a general bias against movies with more prominent female presence. These findings have important implications for the gate-keeping role of review-aggregating platforms in reducing bias against female representation in product ratings.

JEL-Codes: D830, L150, L820, O330.

Keywords: digital platforms, product ratings, critics, crowds, gender bias, movie industry.

Luis Aguiar
University of Zurich / Switzerland
luis.aguiar@business.uzh.ch

October 27, 2024
I thank Diana Andrade, Andreas Bayerl, Grazia Cecere, Pierre-François Darlas, Daniel García, Reinhold Kesler, Michelangelo Rossi, Carmit Segal, and Joel Waldfogel for their feedback as well as participants in the economics seminar at the University of East Anglia, the Center for Competition Policy seminar at the University of East Anglia, the University of Vienna, the University of Zurich, the joint Digital Economics Seminar, Wake Forest University, the 20th ZEW Conference on the Economics of ICTs, and the XXXVI Jornadas de Economía Industrial in Palma de Gran Canaria.

1 Introduction

In markets for cultural goods, review aggregating platforms provide users with product ratings from both professional critics and crowds to guide them around their consumption decisions. Despite these clear benefits, ratings and reviews often incorporate significant biases, particularly regarding the gender of the subject being evaluated. Indeed, a large body of research has reported systemic gender differences and biases against women across various online settings (Bohren et al., 2019; Hinnosaar, 2019; Wu, 2019; Proserpio et al., 2021; Botelho and Gertsberg, 2021; Ederer et al., 2023). In the movie industry, such biases are likely to adversely affect the ratings of films with a more prominent female presence. While both critics and crowds could suffer from such gender biases, there are reasons to believe that these two groups of reviewers may incorporate different information into their evaluations. First, even if they are gender biased, critics have strong incentives to avoid incorporating them into their ratings since doing so would lead to important reputational costs.[1] Second, the groups of critics and crowd reviewers available on review-aggregating platforms differ importantly in their composition. On the one hand, critics are likely to form a relatively homogeneous group of reviewers with similar objectives and incentives. On the other hand, and as reported in recent events, the pool of crowd reviewers is more likely to include individuals with extreme levels of animosity against women and female content. In 2016, the all-female reboot of the movie *Ghostbusters* had received a large share of extremely low ratings from crowd reviewers on IMDb, even before the movie had been theatrically released.[2] Essentially ungated, review aggregating platforms like Rotten Tomatoes allow

[1]A clear illustration of this point can be found in Rotten Tomatoes' definition of the group of critics that the platform considers to be *Top Critics*: "Top Critics exhibit a deep knowledge of film/TV history, and their reviews may also provide valuable cultural context. While their reviews incorporate the lens of their own experience, they also exhibit the ability to remove any biases that may prevent them from serving the audience at-large." See https://www.rottentomatoes.com/critics/top_critics.

[2]See https://screencrush.com/ghostbusters-imdb-what-the/. Such review-bombing behavior, whereby online users assign particular movies with extremely low ratings with the goal of harming their popularity, has often been reported to take place before a movie's release, casting serious doubt on whether reviewers engaging in such behavior had actually seen the movie in question. This led the review-aggregating platform Rotten Tomatoes to eliminate pre-release crowd ratings users in 2019. See https://editorial.rottentomatoes.com/article/making-some-changes/. Similar behavior was reported in 2019 for the woman-lead movie Captain Marvel, in August 2022 following the release of the TV series She-Hulk: Attorney at Law, and in September 2022 following the release of the TV series Ms. Marvel. See, for instance,

anyone to leave ratings anonymously, enabling malicious crowd reviewers to easily express their animosity.

Against this backdrop, the objective of this paper is to address whether crowd-based rating environments provide a gender-neutral source of information relative to professional critics in the context of movie ratings. To do so, I rely on various measures of on-screen female presence - based on the gender composition of a movie's cast - and categorize movies into *male movies* and *female movies*. Under the most restrictive definition, a movie with a female lead actress is categorized as a female movie, while a movie with a male lead actor is categorized as a male movie. I then define the *gender score gap* and the *gender gap in extreme low ratings* as the difference in the rating scores and the share of extreme low ratings given to female and male movies, respectively. Based on this two measures, I explore the following questions. First, how do these gender gaps differ across critics and crowd reviewers? Second, how does the gender of both critics and crowds affect these gender gaps? Third, can review-aggregating platforms adjust the design of their rating systems to reduce gender gaps and potentially screen out biased ratings?

I use data on critic and crowd ratings of 12,657 movies released between the years 2006 and 2021 to test how crowd reviewers affect gender gaps relative to professional critics. More specifically, I rely on what is akin to a difference-in-differences approach by comparing gender gaps (first difference) across crowds and critics (second difference), controlling for movie quality and reviewer characteristics via the inclusion of movie and reviewer fixed effects. Results show that crowd reviewers increase the gender score gap in favor of male movies by between 1.8 and 3 points (on a 10-100 points rating scale) relative to critics.

The analysis further reveals remarkable differences based on the reviewers' gender: relative to male critics, male crowd reviewers exhibit a gender score gap that is over 3 points larger, at the expense of female movies. Consistent with male crowd reviewers showing animosity against prominent female presence in movies, I also find that female movies receive a signifi-

https://www.theverge.com/2019/2/26/18241840/rotten-tomatoes-review-bomb-captain-marvel-star-wars-the-last-jedi, https://www.forbes.com/sites/paultassi/2022/08/18/she-hulk-is-getting-review-bombed-even-harder-than-ms-marvel/?sh=68a9d07251a4, and https://www.forbes.com/sites/paultassi/2022/06/09/ms-marvel-review-bombed-into-being-the-mcus-lowest-scoring-show-on-imdb/?sh=224c11525879.

cantly higher share of extremely low ratings (ratings of 10 points on the 10-100 rating scale) when rated by the crowds. This higher share is exclusively driven by male crowd reviewers. When rated by the latter, female movies receive a share of extremely low ratings that is over 50% higher than that of male movies.

I then explore whether review-aggregating platforms can adjust their design to reduce gender bias in crowd ratings. More specifically, I assess how Rotten Tomatoes' implementation of *verified reviews* – which allows the platform to flag reviews coming from users who can confirm bought a ticket to the movie – affect gender score gaps.[3] Consistent with ill-intentioned reviewers driving the increase in gender gap among crowds, I find that gender gaps increases solely when ratings are left by non-verified crowd reviewers. In other words, male and female movies fare equally well when rated by verified crowd reviewers. Allowing users to verify their reviews can therefore act as a screening mechanism that reduces gender bias among crowd ratings.

Overall, these findings indicate that crowd-based ratings - when left unrestricted - negatively affect movies with a more prominent female presence. These results have important implications for platforms providing consumers with both critic and crowd-based rating information, and they contribute to ongoing discussions by both academics and industry observers about the large under-representation of women in the content, sales, and reviewing of creative products (Smith et al., 2018; Choueiti et al., 2018; Lauzen, 2018; Waldfogel, 2023). First, given that professional critics exhibit no significant gender gap in movie ratings, their availability continues to play an important role for female representation in the movie industry. Second, review-aggregating platforms can rely on relatively simple screening mechanisms - such as the verification of movie ticket purchases by reviewers - to effectively reduce gender bias in crowd-based ratings. Despite the ability of digitization to fully democratize the generation of crowd-based ratings, these findings highlight the importance of the gate-keeping role played by review-aggregating platforms in reducing bias against female representation in the movie industry.

[3]See https://editorial.rottentomatoes.com/article/introducing-verified-audience-score/.

This study relates and contributes to several strands of literature. First, it relates to a large body of work on product ratings in the context of the creative industries (Godes and Mayzlin, 2004; Reinstein and Snyder, 2005; Chevalier and Mayzlin, 2006; Li and Hitt, 2008; Zhu and Zhang, 2010; Godes and Silva, 2012), and more specifically to the literature studying how economic incentives can lead to biases and manipulation in product ratings and reviews (Luca, 2015; Luca and Zervas, 2016; He et al., 2022). I contribute to this literature by showing that such behavior can be driven by non-economic motives and disproportionally affect content with a more prominent female presence. In line with Mayzlin et al. (2014), I show that adjusting platform design can help reduce biases in product ratings. Second, this paper relates to the literature comparing contributions from the crowd and from experts in various contexts (Greenstein and Zhu, 2018; Boudreau, 2019; Greenstein et al., 2021; Sen et al., 2023), including product ratings (Holbrook, 1999; Mollick and Nanda, 2016; Reimers and Waldfogel, 2021). Third, this study contributes to the literature exploring gender bias in the online environment (Lambrecht and Tucker, 2019; Wu, 2019; Bohren et al., 2019; Ederer et al., 2023) and more specifically in the case of product evaluations (Botelho and Gertsberg, 2021; Proserpio et al., 2021; Bayerl et al., 2024). This paper also directly contributes to a more specific set of studies showing important gender inequalities in the creative industries (Goldin and Rouse, 2000; Aguiar et al., 2021; Nagaraj and Ranganathan, 2023; Luo and Zhang, 2022; Leung and Strumpf, 2022; Waldfogel, 2023; Stroube and Waguespack, 2024).[4]

2 Data and Descriptives

2.1 Data

The data for this study come from several sources. I first rely on the Internet Movie Database (IMDb), which provides detailed information on more than half a million movies.[5] Each movie listed on the platform displays an "IMDb rating" which corresponds to the average

[4]While not focusing on gender, Fowdur et al. (2012) explore racial bias in professional critic movie reviews.

[5]See https://www.imdb.com/pressroom/stats/.

rating left by crowd reviewers, on a 1-10 points scale. Moreover, IMDb also provides the distribution of these crowd ratings, i.e. the number of crowd users who left a rating of X points, with $X = 1, 2, 3, \ldots, 10$. For instance, as of September 1 , 2022, 1,461,622 users had left a rating of 10 to the movie *Shawshank Redemption*, while 310,148 users had left a rating of 8.[6] These crowd ratings are further be broken down by demographic characteristics, including gender and age. Finally, IMDb allows to construct variables related to movie-specific characteristics, including the origin of the producers, the release date, the production studio, the script writers, the languages spoken in the movie, and the movie's MPAA rating whenever available.[7]

The second source of data is Rotten Tomatoes, a website that aggregates critic and crowd movie ratings and reviews. For each movie appearing on its site, Rotten Tomatoes features an "All Audience" reviews page which provides a list of crowd-based reviews, together with their rating score, on a 0.5-5 points scale.[8] Unlike IMDb, Rotten Tomatoes unfortunately does not report the gender of the crowd reviewers.[9] However, Rotten Tomatoes identifies each crowd reviewer with a unique identifier, together with the date on which the review was posted, which allows to follow reviewers across movies and over time. Each movie additionally features an "All Critics" page which provides the list of all reviews left by critics, including the original rating score whenever available. Each critic is identified with a unique identifier and by their full name, which allows to infer their perceived gender by matching their first name with databases providing information on names and gender.[10] Rotten Tomatoes also provides a detailed genre classification for each movie.

[6] See `https://www.imdb.com/title/tt0111161/ratings/?ref_=tt_ov_rt`. One limitation of the data is that it only provides the cumulative number of ratings left by demographic group at the time of the website visit and does not provide the time at which the rating was posted. Because individual reviewers are not uniquely identified, it is also not possible to follow individual crowd reviewers across movies.

[7] The MPAA rating is used in the U.S. and its territories to rate a movie's audience suitability based on its content, and consists in 4 distinct categories: NC-17, PG, PG-13, or R. See `https://www.motionpictures.org/film-ratings/`.

[8] See, for instance, `https://www.rottentomatoes.com/m/no_time_to_die_2021/reviews?type=user`.

[9] One could in principle rely on the user names left by individuals next to their ratings to identify their gender. However, the fact that reviewers are free to use any type of name as their identifier renders this approach highly unreliable.

[10] One first source of such information comes from the U.S. Social Security names database. As an additional source, I rely on the website *gender-api.com*, which provides a probability score for each search to determine the gender of a given name and also covers non-US names. See `https://gender-api.com`.

The third source of information is Box-Office Mojo, a website that provides information on movies' domestic as well as international box-office revenues, as well as the number of distinct theaters and countries in which the movie was released.[11]

I use these underlying datasets to construct two main samples that I will use in the analysis. I first retrieve the list of all featured films released between the years 2006 and 2021 that appear on IMDb, including the number of ratings (or votes) left by users for each movie. Given the high concentration of ratings, I further focus on the movies that received at least 100 votes on IMDb.[12] I then match the movies that appear in the IMDb data to Rotten Tomatoes and Box-Office Mojo in order to incorporate information on critics' and Rotten Tomatoes crowd reviewers' ratings as well as box-office revenue.[13] Since the goal of the analysis is to explore how movie ratings are affected by crowds compared to critics, the final sample is restricted to the subset of movies that have been rated by both critics and crowd reviewers. After matching and cleaning the data to discard observations with missing information on movie characteristics, I end up with a final sample of 12,657 movies.[14] For each of these movies, I first obtain the complete list of crowd ratings from IMDb.[15] I then obtain the set of all crowd and critic reviews listed on each movie's "All Audience" and "All Critics" pages from Rotten Tomatoes, respectively. Because the rating scores on IMDb and Rotten Tomatoes – coming from both critics and crowds – differ in their scales, I rescale all ratings to a common 10-100 scale. Appendix B.1 presents more extensive institutional details on IMDb and Rotten Tomatoes and provides a detailed description of how I rescale the ratings of both platforms.

Finally, I use three different measures to quantify female presence in each movie. I start by

[11] See https://www.boxofficemojo.com/.

[12] There is a total of 151,848 featured films released between 2006 to 2021 that are listed on IMDb. Out of these, 44,835 receive at least 100 votes, and this subset of movies accounts for 99.7% of all votes obtained by the 151,848 movies.

[13] Matching movies from Rotten Tomatoes and IMDb is performed using the Open Movie Database, a third party site that provides movie-level information as well as a common identifier across Rotten Tomatoes and IMDb for a large set of movies. See https://www.omdbapi.com/.

[14] This final set of 12,657 movies accounts for 92% of the total number of votes received by the 44,835 movies that received at least 100 votes. The set of movies included in the final sample therefore accounts for the vast majority of the IMDb catalog value according to IMDb votes.

[15] All the data related to ratings by both crowds and critics were collected during the first two weeks of March 2022.

inferring the gender of the top 3 cast members as listed on the movie's IMDb page based on the IMDb profile page of each cast member.[16] I then use a sequence of definitions to classify a movie as female, from more to less restrictive. The first and most restrictive definition considers a movie as female if the lead role of the movie is played by a female actress. The second definition considers a movie as female if at least one of the top 2 members of the cast is female. The third and least restrictive definition considers a movie as female if at least one of the top 3 members of the cast is female. Finally, I also rely on the Bechdel Test, which measures the degree of female presence in a movie based on the interaction between male and female characters. I define a movie as female if it passes the Bechtel Test, which requires that it features at least two women who talk to each other about something other than a man.[17] I also obtain information on the gender of each movie's producers and writers to construct measures of female presence behind the scenes.

2.2 Descriptive Statistics

Panel A of Table 1 provides descriptive statistics on the movies included in the final sample. Out of the 12,657 movies included in the sample, 75% were theatrically released, generating an average of $42.5 million in revenue worldwide. Female cast representation is far from being even, with only 35% of movies presenting an actress as their lead. Two-thirds of movies have at least one actress in their top 2 cast members, and 82% of movies have at least one female in the top 3 members of the cast. Based on the Bechdel Test, about 62% of the movies are female. When rated by critics, movies obtain, on average, 76% of their ratings from male critics. Even though one might perhaps expect a crowd rating environment to level out the contribution of female reviewers, a striking 79% of a movie's ratings comes from men among

[16]On top of providing the full list of a movie's cast, each IMDb movie page also provides the list of the top 3 stars of the film. I therefore define the latter as the top 3 cast members, ranking them according to their order of appearance on the IMDb page.

[17]Information on a movie's Bechdel Test is publicly available at https://bechdeltest.com/ for a set of over 9,000 movies. Out of these, however, only 3,571 can be matched to my list of movies released between 2000 and 2021, accounting for about 28% of the set of movies included in the final sample. Note also that the site provides user-generated information and is therefore operated by moviegoers. Given that individuals who decide to add films to the dataset are likely particularly concerned about female representation in films, it is unlikely that this sample of movies is representative of the full population of movies.

the IMDb crowd, on average.[18] Panel B reports descriptives on reviewer characteristics from both Rotten Tomatoes and IMDb. Critic gender shares are far from being even, with about two thirds of all critics included in the final sample identified as males. Crowd reviewers from both IMDb and Rotten Tomatoes report higher average scores than critics.

Figure 1 presents a more detailed look at the scores assigned to male and female movies by the different types of reviewers. Relying on the most restrictive definition of a female movie (i.e. based on whether the lead actress is female), the figure shows that critics assign similar scores to male and female movies. In contrast, crowd reviewers on both IMDb and Rotten Tomatoes rate female movies 4 points lower than male movies, on average. On their face, these descriptives suggest that crowd-based ratings penalize female movies relative to male movies.

3 Estimation approaches

I now present the empirical approaches used to explore differences in gender gaps across critics and crowds. Given the important differences between the IMDb and Rotten Tomatoes datasets, I describe each approach separately below.

3.1 Approach Using IMDb Crowd Data

As suggested by Figure 1, crowd reviewers exhibit a larger gender score gap compared to critics. Going beyond these descriptives, I estimate the following model to test how crowd-based ratings affect the gender score gap relative to critics:

$$\begin{aligned} Score_{im} = \quad & \alpha + \delta_0 FemaleMovie_m + \delta_1 \left(FemaleMovie_m \times Crowd_i\right) \\ & + \gamma_0 Crowd_i + \beta X_m + \varepsilon_{im}, \end{aligned} \tag{1}$$

[18] This figure is consistent with existing findings relying on similar data (Boyle, 2014; Stroube and Waguespack, 2024; Bayerl et al., 2024). Appendix B.2 further explores differences in the female share of crowd reviewers across male and female movies in more details.

where $Score_{im}$ is the rating score given by reviewer i to movie m, and $FemaleMovie_m$ is a dummy variable equal to 1 if the movie is female, based on the four definitions presented above. $Crowd_i$ is a dummy equal to 1 if reviewer i is from the crowd (as opposed to being a critic), and X_m includes a set of movie characteristics, including an indicator of whether movie m was released in theaters, box-office revenue, number of released countries, and awards won. It also includes a large host of fixed effects, including genre, production studio, MPAA rating, year of release, language, as well as movie origin fixed effects. ε_{im} is an error term.

The coefficient δ_0 corresponds to the movie gender score gap among critics, while the movie gender score gap among crowds is given by $\delta_0 + \delta_1$. The main coefficient of interest is the difference-in-differences coefficient δ_1 which shows the difference in the movie gender score gaps between critics and IMDb crowd reviewers.[19] I can further include movie fixed effects in equation (1) to control for time-invariant unobservable movie quality. While this does not allow for the identification of the gender score gap for the critics and crowds separately, it still allows to identify δ_1, i.e. the difference in the gender score gap across both types of reviewers.

To explore whether crowd-based ratings exhibit a higher level of animosity towards female movies, I further focus on the share of extremely low ratings left by crowd reviewers relative to critics. Defining extremely low ratings as scores of 10 points on the 10-100 rating scale, I can estimate specifications similar to (1) using a dummy equal to 1 for extremely low ratings as a dependent variable. In that case, the coefficient δ_1 captures the difference in the gender gap in extreme low ratings across IMDb crowd and critic reviewers.

3.1.1 The Role of Reviewers' Gender

Given that IMDb provides information on the gender of its crowd reviewers, I can ask whether the latter affects the difference in the gender score gaps across critics and crowds. To explore

[19]Note that this difference-in-differences estimate is rather unconventional in the sense that neither of the differences is in the time dimension. See Mayzlin et al. (2014) for the implementation of a similar approach in the context of the identification of promotional reviews on Tripadvisor and Expedia.

this, I expand equation (1) by interacting the main variables of interest with indicators for the gender of the reviewer:

$$\begin{aligned}
Score_{im} = \quad & \alpha + \theta_0 FemaleMovie_m + \theta_1 \left(FemaleMovie_m \times Crowd_i\right) \\
& + \theta_2 \left(FemaleMovie_m \times Crowd_i \times MaleReviewer_i\right) \\
& + \gamma_0 Crowd_i + \gamma_1 MaleReviewer_i + \gamma_2 \left(FemaleMovie_m \times MaleReviewer_i\right) \\
& + \gamma_3 \left(Crowd_i \times MaleReviewer_i\right) + \beta X_m + \varepsilon_{im}, \qquad (2)
\end{aligned}$$

where the variable $MaleReviewer_i$ is a dummy equal to 1 if reviewer i is male. In this specification, θ_1 corresponds to the difference-in-differences estimate showing the difference in the gender score gap between female critics and IMDb female crowds. Likewise, the coefficient $\theta_1+\theta_2$ corresponds to the difference-in-differences estimate capturing the difference in gender score gaps between male critics and male IMDb crowds. Note that one can again include movie fixed effects in (2) to identify both difference-in-differences estimates of interest while controlling for unobservable movie quality. As above, I can estimate the same specifications as in (2) using a dummy equal to 1 for extremely low ratings as a dependent variable.

3.2 Approach Using Rotten Tomatoes Crowd Data

Unlike IMDb, the Rotten Tomatoes crowd data provide information on the timing of the reviews and allows to follow reviewers both across movies and over time via a unique identifier. I can therefore exploit the panel structure of the data and control for dynamic effects of reviews as well as reviewer characteristics. Previous research has indeed highlighted the importance of dynamic effects in online ratings in a various range of settings.[20] In the context of book reviews, Godes and Silva (2012) showed that ratings are affected by both a temporal process (i.e. depending on the amount of time that a book has been available for review) as well as by a sequential process (i.e. depending on the ordering in which the reviews are

[20]See, for instance, Godes and Mayzlin (2004), Moe and Trusov (2011), Godes and Silva (2012), and Moe and Schweidel (2012).

posted). To account for these two dynamic processes, I start by constructing a variable that measures the position of a review within the sequence of reviews received by a given movie. Because Rotten Tomatoes only provides reviews at the date level without a time stamp, I cannot determine the order of reviews that were posted on the same date. Following Godes and Silva (2012), I deal with this particular issue by giving the same sequence position to all reviews that were posted on the same day. More formally, I construct the variable $Review\_Position_{im}$, which indicates the position of review i within the sequence of reviews posted for movie m up to and including the day on which review i was posted. Define t_i as the day on which review i is posted. Because multiple reviews can be posted on the same day, define R_t as the set of all reviews posted on day t: $R_t = \{i|t_i = t\}$. For each review i of movie m, I then construct the variable $Review\_Position_{im}$ as follows:

$$Review\_Position_{im} = \sum_{\tau<t} |R\tau| + 1,$$

where t is the day on which the review was posted, and $|R\tau|$ is the cardinality of the set $R\tau$. For a given movie m, this approach therefore assigns the same review position to all reviews posted on the same day. More specifically, the position of each review posted on day t is equal to 1 plus the cumulative number of reviews received on days $\tau < t$.[21]

To further account for changes in ratings over time, I construct the variable $Elapsed\_Time_{im}$, which measures the number of days that have elapsed since movie m received its first review, up to the date when review i was posted.[22]

[21] Note that I group both critics and crowd reviews together when constructing the variable $Review\_Position_{im}$. For instance, suppose that movie m obtains its first two critic reviews and its first 3 crowd reviews on June 24. Suppose further that the same movie then receives 3 additional critic reviews and 10 additional crowd reviews on June 25, and 1 additional critic review and 16 crowd reviews on June 26. In that case, the first 5 reviews of June 24 will be assigned values $Review\_Position = 1$, the next 13 reviews of June 25 will be assigned $Review\_Position = 6$, and the next 17 reviews on June 26 will be assigned $Review\_Position = 19$. This approach assumes that a reviewer considers the join ordering of both critics and crowd prevailing reviews when reviewing a movie. An alternative approach would be to assume that a reviewer considers the ordering of critics and crowd reviews separately and to construct critic and crowd-specific orderings. Both approaches lead to quantitatively similar results.

[22] Note that I again construct such variable without distinguishing between critics and crowd reviews. In other words, I calculate the elapsed time between a given review and the first review, regardless of whether the latter was posted by a critic or a crowd reviewer. For instance, suppose that movie m received its first critic review on June 24 and its first crowd review on June 25. Then $Elapsed\_Time_{im}$ will be computed

The panel dimension of the Rotten Tomatoes data further allows to control for two crucial aspects of online reviewing activity. First, I can control for reviewer fixed effects to account for unobservable and time-invariant reviewers' characteristics, for instance the fact that some reviewers (whether critics or crowd) tend to be harsher than others. Second, I can also control for a global trend in rating dynamics via the inclusion of a time trend. More specifically, I include the variable $Review\_Year_{im}$ indicating the year in which review i of movie m was posted.

With the various variables defined above, I can estimate the following model to test how Rotten Tomatoes crowd-based ratings affect the gender score gap relative to critics:

$$\begin{aligned} Score_{im} = \quad & \delta_1 \left(FemaleMovie_m \times Crowd_i \right) + \\ & \lambda Elapsed\_Time_{im} + \phi Review\_Position_{im} + \mu_m + \eta_i + \psi Review\_Year_{im} + \varepsilon_{im}, \end{aligned} \tag{3}$$

where μ_m and η_i are sets of movie and reviewer fixed effects. The main coefficient of interest is the difference-in-differences coefficient δ_1 which captures the difference in the movie gender score gap between critics and Rotten Tomatoes crowd reviewers. Similar to the approach exposed in Section 3.1, I can also explore whether the gender gap in extreme low ratings varies across critics and crowds.

4 Results

This section starts by reporting estimation results based on the IMDb crowd data and then turn to the Rotten Tomatoes data. In both cases, I present the results that rely on the most restrictive definition of a female movie, based on the gender of the lead actor. Throughout the text below, I refer to Appendix B.3 for estimation results that rely on less restrictive

relative to June 24 for both crowd and critic reviews. For instance, both a crowd and critic review posted on June 28 will be assigned $Elapsed\_Time = 4$. Relying on critic and crowd-specific elapsed times lead to quantitatively similar results.

female movie definitions.

4.1 Gender Score Gaps on IMDb

Relying on the IMDb sample, Table 2 presents the results of estimating various specifications of equation (1) using two sets of dependent variables. Focusing on the rating score, specifications (1) indicates that female movies tend to fare slightly better than male movies when reviewed by critics, as they obtain score that are 0.98 points higher on average. As indicated at the bottom of the table, the gender score gap is significantly larger in the crowd-based environment, where female movies obtain ratings that are 2.55 points lower than male movies on average. Specification (2) introduces movie fixed effects to control for time-invariant unobservable movie quality. The difference-in-differences estimate - which captures the difference in the movie gender score gap across critics and crowds - is reduced slightly in magnitude but remains large and significant, indicating that crowds increase the gender score gap by 2.7 points in favor of male movies.

Specifications (3) and (4) of Table 2 follow a similar approach but use a dummy equal to 1 for extreme low ratings (i.e. if the rating assigned is equal to 10 on the 10-100 scale) as the dependent variable. More specifically, I multiply this dummy by 100 in order to interpret the coefficients as differences in percentage points and facilitate the readability of the results. Specification (3) indicates that while critics tend to assign a slightly lower share of extreme low ratings to female movies (by 0.2 percentage points), the share of extremely low ratings assigned to female movies (relative to male movies) is 1.1 percentage points higher among the crowds. Even after controlling for movie fixed effects, crowds increase the gender gap in extreme low ratings by 0.93 percentage points at the expense of female movies. These results again indicate that female movies fare worse than their male counterpart when being rated by the crowds.[23]

[23]Tables A1 and A2 in Appendix B.3 show the results of analogous estimations when relying on less restrictive definitions of a female movie. The differences reported in Table 2 remain significant although the corresponding coefficients decrease in magnitude, perhaps unsurprisingly, as less restrictive measures of movie gender are used.

4.1.1 The Role of Reviewers' Gender

Table 3 shows the results of estimating equation (2) again using the same two dependent variables. To facilitate the reading of the results, the coefficients corresponding to the movie gender score gaps in the four rating environments (female-critics, female-crowds, male-critics, and male-crowds) are plotted in Figure 2. The left panel shows the corresponding coefficients of interest using the rating score as a dependent variable. Focusing on female critics, the first estimate of the panel shows a positive and statistically significant gender score gap, indicating that female movies fare better than male movies within that group of reviewers, by about 1.9 points. When rated by female crowd reviewers, however, female movies fare significantly worse than male movies, by about 1.2 points. Male critics tend to assign slightly better scores to female than to male movies, although this difference is small at 0.7 points and barely significant. Most striking is the magnitude of the gender score gap when movies are rated by male crowds. In this case, and relative to male movies, female movies are assigned scores that are lower by over 3.2 points.

The right panel of Figure 2 reports the corresponding estimates when focusing on extreme low ratings. These make clear that the increase in the share of extremely low ratings associated with crowd reviewers - reported in Table 2 - is exclusively driven by male crowd reviewers. When rated by male crowds, female movies receive a share of extremely low ratings that is 1.17 percentage point higher than male movies, corresponding to a more than 50% difference.[24] When rated by male critics and female crowds, however, both male and female movies receive similar shares of extremely low ratings. Female movies nevertheless receive a significantly lower share of extremely low ratings than male movies when rated by female critics (by about 0.5 percentage points). Overall, these results once again highlight that a) relative to critic ratings, crowd ratings lead to a higher gender gap at the expense of female movies, and b) this increase in inequality is significantly larger for male than female reviewers.[25]

[24]Within the sample, the share of extreme low ratings for male movies reaches 2.17% when rated by male crowd reviewers.

[25]For robustness, Figures A1 and A2 in Appendix B.3 report the corresponding estimates when using alternative definitions of female movies.

Figure 3 presents the difference-in-differences estimates when controlling for movie fixed-effects. For comparison, each panel also reports the estimates when combining all reviewers together, corresponding to the difference-in-differences estimates presented in columns (2) and (4) of Table 2. The female coefficient in the left panel of Figure 3 reports the difference-in-differences estimate for female reviewers, indicating an increase of about 2.16 points in the movie gender score gap when comparing female-critics and female-crowds.[26] When focusing on male reviewers, the gender score gap again increases to a much larger extent, by as much as 3.2 points in favor of male movies. Comparing the gender-specific estimates with the ones that combine all reviewers together (the top estimate in each panel) clearly indicates that most of the increase in the gender score gap reported in Table 2 is driven by male crowd reviewers. The right panel of Figure 3 presents similar results when focusing on the gender gap in extreme low ratings. Once again, most of the increase in the gender gap in extreme low ratings reported in Table 2 is driven by male crowd reviewers.[27]

4.2 Gender Gaps on Rotten Tomatoes

I now turn to the estimation results based on the Rotten Tomatoes data. As explained above, such data allow us to employ a richer empirical approach by exploiting the panel dimension of the data.

Table 4 presents the results of estimating equation (3), with specifications (1)-(4) focusing on the rating score as a dependent variable and specifications (5)-(8) focusing on extreme low ratings. Looking at the rating scores, it is interesting to note that specifications (1) and (2) show very close estimates compared to the IMDb crowd data (see Table 2). In other words, the crowd-based gender score gaps are very similar on both Rotten Tomatoes and IMDb. Specification (3) additionally controls for reviewer fixed effects as well as both the

[26] This corresponds to the estimate of coefficient θ_1 in Table 3. As indicated in the left panel of Figure 2, this increase is mostly driven by the fact that while female movies receive lower ratings than male movies when rated by female crowds, they also tend to receive better ratings than male movies when rated by female critics.

[27] Figures A3 and A4 in Appendix B.3 report the corresponding estimates when using alternative definitions of female movies.

sequential and temporal dynamics of ratings. Including such controls reduces the difference-in-differences estimate, but the latter still remains significant both in magnitude and statistically.[28] As indicated by the coefficients on the $Elapsed\_Time_{imt}$ and $Review\_Position_{imt}$ variables, the ratings appear to decline in both the time since the first review and the sequential order in which they are posted. Specification (4) further includes a time trend, and the gender score gap remains significantly larger under crowd reviews relative to critics, by 1.8 points.[29]

Regarding extreme low ratings, specifications (5) and (6) indicate that Rotten Tomatoes crowd reviewers report larger gaps relative to the IMDb crowd. Controlling for movie fixed effects, the gender gap in extreme low ratings among crowd reviewers increases by about 1.6 percentage points relative to critics. This gap remains significant after controlling for reviewer fixed effects and temporal dynamics, although it decreases in magnitude. As reported in specification (8), crowd reviewers increase the gender gap in extreme low ratings by 0.91 percentage points relative to critics.[30]

4.3 Robustness

This section explores the robustness of these results to potentially unobservable differences across male and female movies and to alternative information measures provided in reviews.

[28]Note that controlling for reviewer fixed effects leads to the removal of a substantial number of crowd reviewers from the sample. This is because 59% of crowd reviewers (2,751,187 out of 4,662,334) post only a single review throughout the sample period. To assess how these reviewers affect the gender score gap, I re-estimate specification (2) by excluding these single reviewers from the sample. The resulting difference-in-differences coefficient becomes slightly lower in magnitude, at -2.15 (p-val = 0.000). This indicates that the large set of single reviewers leads to a slightly larger gender score gap within the crowds. In that sense, the specifications including reviewer fixed effects are likely to provide a conservative measure of the difference-in-differences estimates.

[29]Note that the coefficient on the time trend is negative, indicating that movie rating scores on Rotten Tomatoes globally tend to decrease over the period of time 2006-2021. Controlling for review year fixed effects as opposed to a linear trend leads to quantitatively similar results. Moreover, including the time trend variable flips the sign of the coefficient on the $Elapsed\_Time_{imt}$ variable. Ratings therefore tend to increase with the elapsed time since the first review once controlling for the common underlying trend in rating scores. It is worth noting that these results are identical to the ones reported by Godes and Silva (2012) in their study of book ratings on Amazon.

[30]Table A3 in Appendix B.3 presents the results of estimating specification (4) and (8) using alternative definitions of female movies. Results show that even when based on less restrictive definitions, female movies fare worse than male movies when rated by the crowds.

It additionally considers important social dynamic aspects of ratings and reviews.

4.3.1 Unobservable Differences in Movie Content

One potential concern when comparing movies with female and male leads is that these two groups of movies could differ in their content, which could be correlated with gender. For instance, it could be that female movies also exhibit more progressive politics. To check that the results presented above are truly driven by differences in on-screen female presence rather than underlying differences in content, I classify movies according to the gender composition of their script writers, which are likely to influence the content of a movie yet are less likely to be observed by reviewers.[31] I then perform similar estimations as above, but within the subset of female-written movies and male-written movies separately. In other words, I ask whether the increase in gender gaps between critics and crowds is found among both male-written movies and female-written movies. I perform similar estimations based on the gender composition of the movie producers. Appendix B.4 reports the results of this exercise and shows similar difference-in-differences estimates across all subsamples, providing support to the fact that the increase in the gender gaps documented above are driven by differences in female presence among the top cast rather than differences in movie content.

4.3.2 Sentiment Analysis

Given that Rotten Tomatoes provides the text of the reviews left by both critics and crowds along with their rating score, I can construct alternative rating measures based on a text sentiment analysis. I rely on three rule-based analyzers to do so. I first rely on VADER (Valence Aware Dictionary for sEntiment Reasoning) and TextBlob's sentiment analyzer, two popular tools that provide measures of the sentiment of a given text, ranging from -1 (most negative sentiment) to 1 (more positive sentiment). I additionally use the LIWC

[31] I thank the Associate Editor for this suggestion. Note that while one cannot rule out that the gender composition of script writers could be observed by reviewers, it is reasonable to assume that it is much less likely to be observed than the gender composition of the cast.

(Linguistic Inquiry and Word Count) tool to obtain a measure of the percentage of negative-tone words (e.g. “bad,” “hate,” “wrong” etc.) out of a given review (Pennebaker et al., 2001; Boyd et al., 2022). This measures accordingly ranges from 0 (no words with negative tone) to 100 (all words with negative tone). The results of estimating equation (3) relying on the three sentiment score measures are reported in Appendix B.5 and confirm the robustness of the main results as crowd reviewers consistently increase gender gaps in the movie reviews' sentiment relative to critics.

4.3.3 Social Dynamics

Social dynamics have been shown to importantly affect online product ratings (Godes and Mayzlin, 2004; Moe and Trusov, 2011; Godes and Silva, 2012; Moe and Schweidel, 2012). Beyond controlling for the sequential and temporal dynamics of ratings in equation (3), it is worth exploring whether the gender gaps documented above materialize at different points in time within a movie's lifetime, for at least two reasons. First, one may wonder whether earlier reviews exhibit more animosity against female movies. Second, it is important to take into account that reviewers might be influenced by other users' opinions. In particular, reviewers may engage in herding and follow the ratings of previous reviewers (Banerjee, 1992; Bikhchandani et al., 1992). If this is the case, then only the early reviews would contain relevant information about a given movie. Finally, focusing on different periods of a movie's life also accounts for the fact that early reviews are more likely to matter relative to reviews posted later on, especially given the very large number of posted ratings (Godes and Mayzlin, 2004). I explore these dynamic aspects in Appendix B.6. While the results are consistent with herding behavior, they also show that the main results remain robust even after controlling for such social dynamics. They also reveal that the differences in the gender score gap across crowds and critics realize both in the early and late life of movies.

4.4 Gender Bias and Platform Design Choices

The findings above that show crowd-based ratings negatively affect female movies. While these results might be driven by a general gender bias among crowd reviewers, they may also result from the ungated nature of review-aggregating platforms, which allow explicitly malicious reviewers to easily express their animosity. Platform design choices about who can post ratings can indeed importantly affect the composition of the crowd reviewers' pool, and in particular incentivize the participation of hostile reviewers. As reported in recent events, this feature is likely to disproportionately affect female movies and can, at the extreme, lead some platform users to easily engage in review bombing against movies with a strong female presence.[32] The results showing that female movies are assigned a significantly higher share of extreme low ratings when rated by crowd reviewers – in Table 4 – provide indicative evidence of such a mechanism being at play.

Given the above, I explore whether platforms can adjust their design to reduce gender bias among crowd ratings. To do so, I exploit a change implemented by Rotten Tomatoes in May of 2019, where the platform introduced *verified reviews* in an effort to increase their reliability. Verified reviews are reviews that are posted by reviewers who were able to demonstrate that they saw the movie they reviewed.[33] For each movie listed on the platform, Rotten Tomatoes then features the list of verified reviews separately and prominently.[34] Such verification has at least two relevant implications in this context. First, verified reviews allow to screen out malicious reviewers whose goal is simply to bomb a given movie's ratings without having watched it. Second, because the cost of posting a verified review is higher, verified reviews can also attract more dedicated crowd users. Given that the platform puts more weight

[32]See, for instance, https://www.forbes.com/sites/paultassi/2022/08/18/she-hulk-is-getting-review-bombed-even-harder-than-ms-marvel/?sh=68a9d07251a4.

[33]To verify their review, users must upload a copy of the receipt of the ticket they purchased. At the time of this writing, Rotten Tomatoes can only verify reviews of users who purchased their ticket online via the the ticketing company Fandango (which owns Rotten Tomatoes) using the same email address than the one used for their Rotten Tomatoes account. This feature implies that verified users may consist of users with socio-economic backgrounds that are not representative of all crowd reviewers. I discuss the implications and limitations of such verification system in Section 5.

[34]See, for instance, https://www.rottentomatoes.com/m/no_time_to_die_2021/reviews?type=verified_audience. Additionally, the default Tomatomater score provided by Rotten Tomatoes (which indicates the share of all audience reviews with 3.5 stars out of 5 ore more) is based on verified reviews only.

on verified reviews and exposes them more prominently, reviewers may also consider them as having a higher status or relevance. In other words, verified reviews may bring in a higher sense of responsibility for reviewers and trigger the posting of reviews that rely less on factors unrelated to quality, like gender.[35] Taken together, these considerations highlight the potential screening role that review verification may play in reducing gender bias within crowd reviewers.

To explore this, I start by descriptively comparing verified and non-verified reviewers.[36] Because Rotten Tomatoes introduced verified reviews on May 23, 2019, I can only rely on the subset of movies and ratings released after that date to perform any comparisons. This leads to a smaller sample that includes a total of 588 movies. Out of all crowd-based ratings observed in this restricted sample, 53% are verified. Table 5 reports differences in the share of extreme low ratings left by verified and non-verified crowd reviewers. The figures indeed indicate that verified users form a selected group of users that differs importantly from non-verified reviewers. Panel A shows how verified reviewers tend to post a substantially lower share of extremely low ratings and reviews. Consistent with the arguments laid out above, panel C indicates that in the case of female movies, the share of extreme low ratings posted by non-verified reviewers is over 9 percentage points higher than the share of extreme low ratings posted by verified reviewers.

4.4.1 Gender Gaps among Verified and Non-Verified Crowds

Table 6 reports the results of estimating equation (3) using the Rotten Tomatoes crowd data on the restricted sample of movies released after May 2019. Specification (2) presents a slight variation of specification (1) to control for reviewer characteristics. Instead of including reviewer fixed effects – which lead to a loss of a large share of crowd reviewers who only leave

[35] In the context of restaurant ratings, Botelho and Gertsberg (2021) show that status awards can indeed have a disciplining effect on evaluators, leading to less gender-biased reviews. While verified reviews can naturally not be considered as awards, their increased visibility and salience may still generate similar effects.

[36] Note that a same user may verify their review for a given movie but not for another. I define verified reviewers as reviewers who have posted at least one verified review during the sample period. This definition has no impact on the results presented below.

a single review – I construct the variable $Average\_Score\_Others_{im}$, which gives the average score of all ratings given by reviewer i to the movies other than m within the whole sample.[37] Controlling for reviewer characteristics in such a way leads to a slightly larger difference-in-differences coefficient compared to directly controlling for reviewer fixed effects.[38]

To test for differences in the gender score gap across verified and non-verified crowd ratings, I further interact the variables of interest of equation (3) with indicators for verified and non-verified crowd reviewers. Columns (3) and (4) show that the increase in the gender score gap identified in column (2) is exclusively driven by non-verified crowd reviewers. While the gender score gap is indistinguishable across critics and verified crowd reviewers both in magnitude and statistically, it is substantially larger for non-verified crowd reviewers compared to critics, by over 4 points, as indicated in column (4).

Using an indicator for extreme low ratings as the dependent variable, columns (5) and (6) show that the gender gap in extreme low ratings increases by about 1.5 percentage points when ratings are assigned by crowd reviewers. Columns (7) and (8) again show that this increase in the gender gap is exclusively driven by non-verified crowd reviewers. Among the latter group, the gender gap in extreme low ratings increases by as much as 2.5 percentage points relative to the critics rating environment.[39]

Overall, these results indicate that allowing crowd users to verify their reviews can serve as a screening mechanism to effectively reduce gender bias in crowd ratings. Moreover, this evidence suggests that the negative impact of crowd ratings on female movies is mainly driven by a particularly hostile subset of reviewers who are less likely to be verified.[40] I

[37]Note that a given crowd reviewer may only have posted a single ratings in the restricted 2019-2021 sample but may have posted earlier reviews before 2019. Using the variable $Average\_Score\_Others_{im}$ instead of reviewer fixed effects therefore allows to keep such reviewers in the 2019-2021 sample.

[38]Consistent with the results discussed in footnote 28, single crowd reviewers therefore tend to increase, if anything, the gender score gap among crowds.

[39]Tables A4 and A5 in Appendix B.3 present the estimation results when relying on alternative female movie definitions. Additionally, Tables A6, A7, and A8 in Appendix B.5 present the corresponding estimates when relying on the review text sentiment score measures. All show results that are consistent with the ones presented in Table 6, again providing support to the robustness of the main results presented above.

[40]Note that this does not imply that all non-verified reviewers are malicious. In fact, it is very likely that an important subset of the non-verified reviews come from genuine reviewers who were simply not able to prove they had seen the movie, for a variety of reasons. There are indeed several reasons why a given review could not be verified by Rotten Tomatoes. A user who purchased their ticket directly at a movie theater,

discuss the implications and limitations of such design below.

5 Discussion and Conclusion

This paper explores whether crowd-based information sources provide gender-neutral product ratings relative to critics. The movie industry, where online animosity against prominent female representation in films has been repeatedly reported, provides an opportunity to empirically test this hypothesis. Using a sample of 12,6567 movies, I show that gender score gaps as well as the gender gaps in extreme low ratings increase significantly with crowd-based ratings and at the expense of female movies. The results also provide evidence suggesting that review verification can act as a screening mechanism that reduces gender bias among crowd ratings. Put differently, review-aggregating platforms can play an important gatekeeping role in reducing bias against female representation through the design of their crowd rating systems. Because the type of evaluations under consideration have become ubiquitous with digitization, these finding also have applicability beyond the movie industry. Indeed, professional and crowd-based ratings are available in a large array of settings such as books, video games, music, and restaurants, among others.

These results naturally have more specific implications for review-aggregating platforms in the movie industry. While concerns about gender balance in the pool of professional critics are often raised (Lauzen, 2018; Choueiti et al., 2018), these findings suggest that crowd reviewers may play an even more important role for on-screen female representation in the movie industry. In fact, the availability of professional critics continues to play an essential function given that they exhibit no significant gender gap in movie ratings. These results also highlight the various trade-offs faced by review-aggregating platforms when designing their rating systems. While crowd ratings suffer from gender bias, they can still provide relevant information on virtually all movies, including the ones that would otherwise not have been rated by critics. Moreover, even if review aggregating platforms can engage in

for instance, would not be able to have their review verified.

review verification to reduce gender bias, such policy comes at a cost and is not without limitations. First, implementing a verification system for all movies is challenging given that providing proof of purchase or movie visioning is often difficult. The current verification system implemented by Rotten Tomatoes, for instance, only allows verification of movies that are theatrically released in the US. Given that many movies are never released in theaters, and considering the important growth in movies released via SVOD platforms, such system only benefits a share of all extant movies. Second, even within the set of movies whose reviews can be verified, platforms need to account for the fact that some unverified reviews may still be genuine and informative. In particular, one should consider the fact that some viewers may simply not be able to verify their purchase, which may lead to a pool of verified reviewers that is not representative of the overall population of movie viewers. Implementing a verification system therefore requires trading-off a larger number of unverified reviews for a lower number of verified ones. From that perspective, Rotten Tomatoes' policy of flagging verified and unverified reviews (and therefore still allowing platform users to access unverified reviews) seems useful, although unverified reviews may still likely suffer from gender bias. The inclusion of such aspects in the design of rating systems suggests fruitful avenues for future research.

References

Aguiar, L., J. Waldfogel, and S. Waldfogel (2021): "Playlisting favorites: Measuring platform bias in the music industry," *International Journal of Industrial Organization*, 78, 102765.

Banerjee, A. V. (1992): "A simple model of herd behavior," *The quarterly journal of economics*, 107, 797–817.

Bayerl, A., Y. Dover, H. Riemer, and D. Shapira (2024): "Gender rating gap in online reviews," *Nature Human Behavior*.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, 100, 992–1026.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The dynamics of discrimination: Theory and evidence," *American Economic Review*, 109, 3395–3436.

Botelho, T. L. and M. Gertsberg (2021): "The Disciplining Effect of Status: Evaluator Status Awards and Observed Gender Bias in Evaluations," *Management Science*.

Boudreau, K. (2019): "'Crowds' of Amateurs & Professional Entrepreneurs in Marketplaces," *Available at SSRN 2988308*.

Boyd, R. L., A. Ashokkumar, S. Seraj, and J. W. Pennebaker (2022): "The development and psychometric properties of LIWC-22," *Austin, TX: University of Texas at Austin*, 1–47.

Boyle, K. (2014): "Gender, comedy and reviewing culture on the Internet Movie Database," *Participations*, 11, 31–49.

Charmes, J. (2019): "The Unpaid Care Work and the Labour Market. An analysis of time use data based on the latest World Compilation of Time-use Surveys," *International Labour Office*.

Chevalier, J. A. and D. Mayzlin (2006): "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, 43, 345–354.

Choueiti, M., S. L. Smith, K. Pieper, and A. Case (2018): "Critic's choice? Gender and race/ethnicity of film reviewers across 100 top films of 2017," *USC Annenberg Inclusion Initiative*.

Ederer, F., P. Goldsmith-Pinkham, and K. Jensen (2023): "Anonymity and identity online," *Working Paper*.

Fowdur, L., V. Kadiyali, and J. Prince (2012): "Racial bias in expert quality assessment: A study of newspaper movie reviews," *Journal of Economic Behavior & Organization*, 84, 292–307.

Godes, D. and D. Mayzlin (2004): "Using online conversations to study word-of-mouth communication," *Marketing science*, 23, 545–560.

Godes, D. and J. C. Silva (2012): "Sequential and temporal dynamics of online opinion," *Marketing Science*, 31, 448–473.

Goldin, C. and C. Rouse (2000): "Orchestrating impartiality: The impact of" blind" auditions on female musicians," *American Economic Review*, 90, 715–741.

Greenstein, S., G. Gu, and F. Zhu (2021): "Ideology and Composition Among an Online Crowd: Evidence from Wikipedians," *Management Science*, 67, 3067–3086.

Greenstein, S. and F. Zhu (2018): "Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopædia Britannica and Wikipedia," *MIS Quarterly*.

He, S., B. Hollenbeck, and D. Proserpio (2022): "The market for fake reviews," *Marketing Science*.

Hinnosaar, M. (2019): "Gender inequality in new media: Evidence from Wikipedia," *Journal of Economic Behavior & Organization*, 163, 262–276.

Holbrook, M. B. (1999): "Popular appeal versus expert judgments of motion pictures," *Journal of consumer research*, 26, 144–155.

Hutto, C. and E. Gilbert (2014): "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 216–225.

Lambrecht, A. and C. Tucker (2019): "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management science*, 65, 2966–2981.

Lauzen, M. M. (2018): "Thumbs down 2019: Film critics and gender, and why it matters," *Report, Center for the Study of Women in Television and Film, San Diego State University*.

Leung, T. C. and K. S. Strumpf (2022): "Discrimination in the Publishing Industry?" *Available at SSRN 3660289*.

Li, X. and L. M. Hitt (2008): "Self-selection and information role of online product reviews," *Information systems research*, 19, 456–474.

Luca, M. (2015): "User-generated content and social media," in *Handbook of media Economics*, Elsevier, vol. 1, 563–592.

Luca, M. and G. Zervas (2016): "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management Science*, 62, 3412–3427.

Luo, H. and L. Zhang (2022): "Scandal, social movement, and change: Evidence from# MeToo in Hollywood," *Management Science*, 68, 1278–1296.

Mayzlin, D., Y. Dover, and J. Chevalier (2014): "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review*, 104, 2421–55.

Moe, W. W. and D. A. Schweidel (2012): "Online product opinions: Incidence, evaluation, and evolution," *Marketing Science*, 31, 372–386.

Moe, W. W. and M. Trusov (2011): "The value of social dynamics in online product ratings forums," *Journal of marketing research*, 48, 444–456.

Mollick, E. and R. Nanda (2016): "Wisdom or madness? Comparing crowds with expert evaluation in funding the arts," *Management science*, 62, 1533–1553.

Nagaraj, A. and A. Ranganathan (2023): "Cutting Through the (Digital) Clutter: Technological Change and Careers of Men and Women in Cultural Markets," *Working Paper*.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001): "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

Proserpio, D., I. Troncoso, and F. Valsesia (2021): "Does gender matter? The effect of management responses on reviewing behavior," *Marketing Science*, 40, 1199–1213.

Reimers, I. and J. Waldfogel (2021): "Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings," *American Economic Review*, 111, 1944–71.

Reinstein, D. A. and C. M. Snyder (2005): "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics," *The Journal of Industrial Economics*, 53, 27–51.

Sen, A., T. Grad, P. Ferreira, and J. Claussen (2023): "(How) does user-generated content impact content generated by professionals? Evidence from local news," *Management Science*.

Smith, S. L., M. Choueiti, K. M. Pieper, A. Case, and A. Choi (2018): *Inequality in 1,100 popular films: Examining portrayals of gender, race/ethnicity, LGBT & disability from 2007 to 2017*, USC Annenberg Inclusion Initiative.

Stroube, B. K. and D. M. Waguespack (2024): "Status and consensus: Heterogeneity in audience evaluations of female-versus male-lead films," *Strategic Management Journal*.

Waldfogel, J. (2023): "The Welfare Effect of Gender-Inclusive Intellectual Property Creation: Evidence from Books," Tech. rep., National Bureau of Economic Research.

Wu, A. H. (2019): "Gender Bias in Rumors among Professionals: An Identity-based Interpretation," *Review of Economics and Statistics*, 1–40.

Zhu, F. and X. Zhang (2010): "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of Marketing*, 74, 133–148.

A Figures and Tables

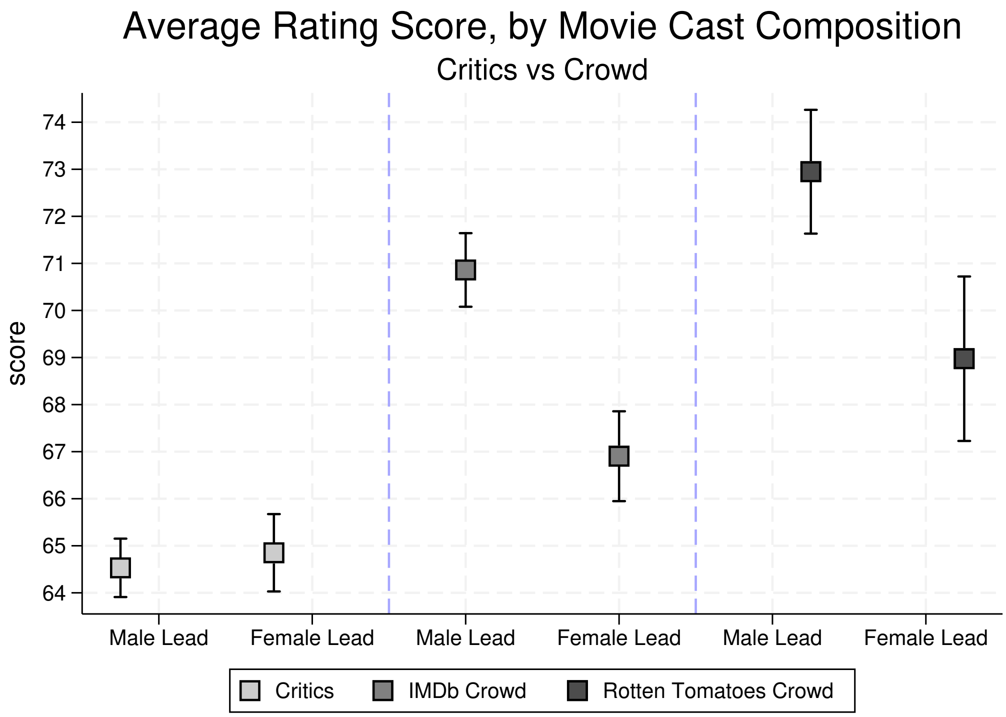
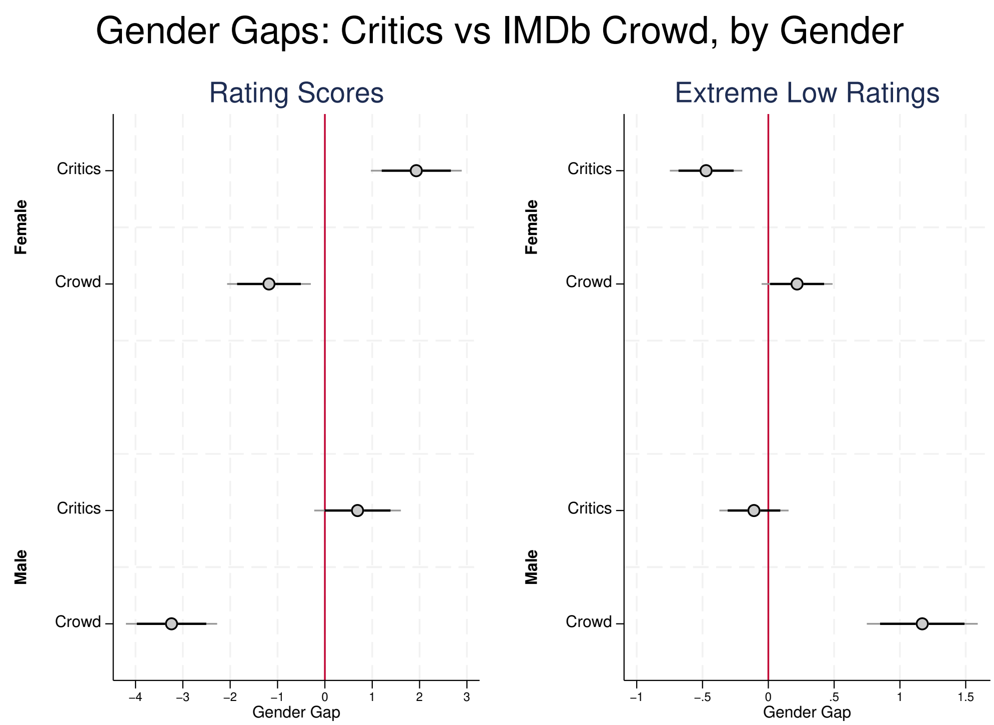
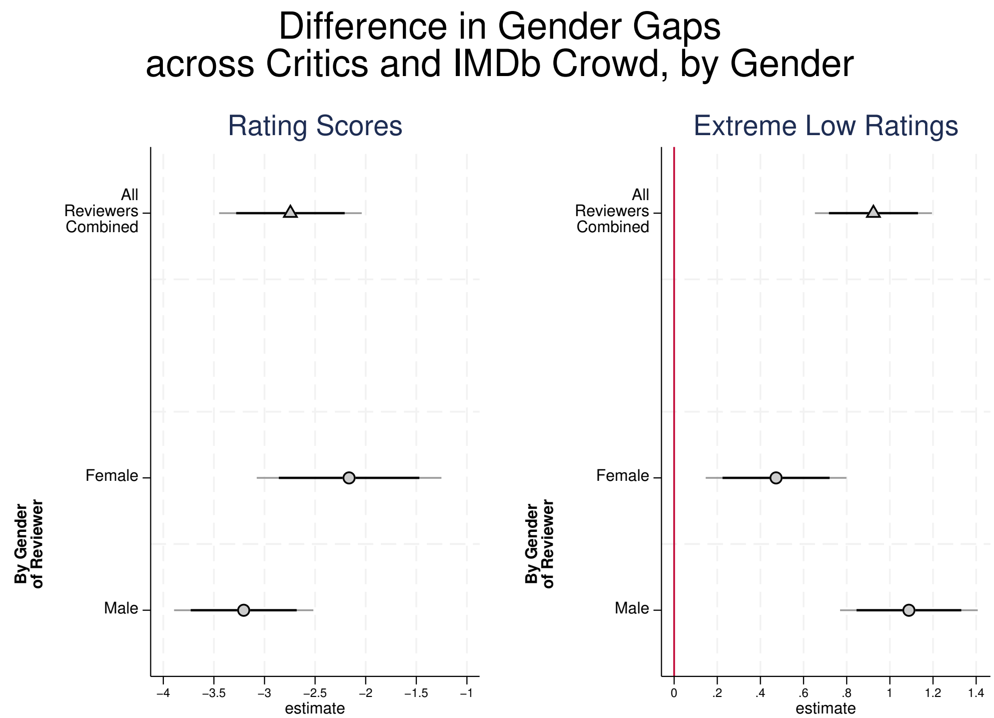


Figure 1: Average Rating Score for Male and Female Movies, Based on Gender of Lead Actor.



Note: The left panel presents estimates of the gender score gap, which is defined as the difference in rating score given to female and male movies. The right panel presents estimates of the gender gap in extreme low ratings (ratings of 10 out of 100), which is defined as the difference in the share of extremely low ratings given to female and male movies. Each coefficient represents the gender gap for the corresponding type of reviewer. A movie is considered female if the top actress is female. For each panel, all estimates come from a single regression as described in the text. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure 2: Gender Score Gap, by Reviewer Type and Gender.



Note: The left panel presents difference-in-differences estimates defined as the difference in the gender score gaps across critics and crowd reviewers. For each type of reviewer, the gender score gap is itself defined as the difference score given to female and male movies. The right panel presents difference-in-differences estimates defined as the difference in the gender gap in extreme low ratings across critics and crowd reviewers. For each reviewer, the gender gap in extreme low ratings is itself defined as the difference in the share of extreme low ratings (ratings of 10 out of 100) given to female and male movies. For comparison, the difference-in-differences estimate for all reviewers combined - corresponding to the estimates from Table 2 - is displayed in each panel. A movie is considered female if the top actress is female. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure 3: Difference-in-Differences Estimates, by Reviewers' Gender.

Table 1: Descriptive Statistics.[†]

	Mean	Std. Dev.	Min	Max	N
A. Movie Characteristics					
Movie Released in Theaters	0.75	0.44	0	1	12,657
Worldwide Box-Office Revenue	42.57	138.01	0	2,798	9,449
Number of Critic Ratings	39.95	55.46	1	361	12,657
Number of IMDb Crowd Ratings	26,615.69	74,636.05	29	1,729,270	12,657
Number of Rotten Tomatoes Crowd Ratings	1,427.86	7,053.71	1	278,844	12,657
Female lead	0.35	0.48	0	1	12,657
Female in Top 2 Cast	0.66	0.47	0	1	12,657
Female in Top 3 Cast	0.82	0.39	0	1	12,657
Passes Bechdel Test	0.62	0.49	0	1	3,571
Share of Male Reviews (Critics)	0.76	0.22	0	1	12,657
Share of Male Reviews (IMDb Crowd)	0.79	0.11	0	1	12,657
B. Reviewer Characteristics					
Rotten Tomatoes Critics					
Male Critic	0.64	0.48	0	1	5,727
Number of Reviews per Critic	88.29	258.39	1	4,770	5,727
Rating Score	64.64	19.81	10	100	505,653
Rotten Tomatoes Crowd					
Number of Reviews per Reviewer	3.88	16.69	1	2,394	4,662,334
Rating Score	72.06	24.71	10	100	18,072,418
IMDb Crowd					
Rating Score	69.94	19.93	10	100	336,874,744

[†] Box office revenue is measured in million USD. Movies' cast composition are based on the cast members appearing on IMDb.

Table 2: Gender Gaps between Critics and IMDb Crowds. †

Dependent Variable:	Rating Score		Extreme Low Rating	
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.
Female Movie	0.986***		-0.213**	
	(0.352)		(0.097)	
Female Movie × Crowd Reviewer	-3.537***	-2.744***	1.124***	0.924***
	(0.367)	(0.273)	(0.137)	(0.105)
Crowd Reviewer	4.223***	2.362***	1.020***	1.546***
	(0.194)	(0.158)	(0.059)	(0.052)
Worldwide Box Office Revenue	0.005***		0.000	
	(0.001)		(0.000)	
Movie Released in Theaters	1.688***		-0.889***	
	(0.565)		(0.316)	
Number of Release Countries	0.069***		-0.031***	
	(0.010)		(0.003)	
Number of Awards Won	0.073***		-0.004***	
	(0.008)		(0.002)	
Genre Fixed Effects	✓	✗	✓	✗
Production Studio Fixed Effects	✓	✗	✓	✗
MPAA Rating Fixed Effects	✓	✗	✓	✗
Year of Release Fixed Effects	✓	✗	✓	✗
Origin Fixed Effects	✓	✗	✓	✗
Language Fixed Effects	✓	✗	✓	✗
Movie Fixed Effects	✗	✓	✗	✓
Gender Gap from Crowd	-2.550		0.912	
p-value	0.000		0.000	
R^2	0.091	0.203	0.011	0.071
Number of Observations	337380397	337380397	337380397	337380397

† A movie is considered female if the top actress is female. Specifications (1) and (2) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (3) and (4) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Box office revenue is measured in million USD. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 3: Gender Gaps between Critics and IMDb Crowds, by Gender. †

Dependent Variable:	Rating Score		Extreme Low Rating	
	(1)	(2)	(3)	(4)
	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.
Female Movie	1.931***		-0.473***	
	(0.372)		(0.107)	
Female Movie × Crowd	-3.112***	-2.164***	0.691***	0.472***
	(0.415)	(0.354)	(0.124)	(0.127)
Female Movie × Crowd × Male Reviewer	-0.817***	-1.041***	0.588***	0.616***
	(0.268)	(0.248)	(0.134)	(0.150)
Crowd	5.173***	3.286***	1.317***	1.886***
	(0.239)	(0.202)	(0.072)	(0.069)
Male Reviewer	-0.601***	-0.600***	0.410***	0.454***
	(0.096)	(0.088)	(0.053)	(0.054)
Female Movie × Male Reviewer	-1.241***	-1.222***	0.364***	0.395***
	(0.174)	(0.158)	(0.085)	(0.089)
Crowd × Male Reviewer	-1.154***	-1.117***	-0.366***	-0.421***
	(0.138)	(0.123)	(0.061)	(0.062)
Worldwide Box Office Revenue	0.005***		0.000	
	(0.001)		(0.000)	
Movie Released in Theaters	1.752***		-0.903***	
	(0.565)		(0.316)	
Number of Release Countries	0.069***		-0.031***	
	(0.010)		(0.003)	
Number of Awards Won	0.073***		-0.005***	
	(0.008)		(0.002)	
Movie Fixed Effects	✗	✓	✗	✓
Difference-in-Differences for Male Reviewers	-3.929	-3.206	1.279	1.088
p-value	0.000	0.000	0.000	0.000
R^2	0.094	0.206	0.011	0.071
Number of Observations	337380397	337380397	337380397	337380397

† A movie is considered female if the top actress is female. Specifications (1) and (2) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (3) and (4) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Specifications that do not include movie fixed effects include genre, production studio, MPAA movie rating, year of release, origin, and movie language fixed effects. Box office revenue is measured in million USD. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 4: Gender Gaps between Critics and Rotten Tomatoes Crowds. †

Dependent Variable:	Rating Score				Extreme Low Rating			
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie	0.421				-0.620***			
	(0.392)				(0.147)			
Female Movie × Crowd Reviewer	-3.153***	-2.263***	-1.781***	-1.785***	2.054***	1.638***	0.905***	0.906***
	(0.628)	(0.409)	(0.297)	(0.297)	(0.319)	(0.204)	(0.125)	(0.125)
Crowd Reviewer	6.517***	5.167***			3.318***	3.104***		
	(0.318)	(0.213)			(0.145)	(0.089)		
Worldwide Box Office Revenue	0.004***				-0.001***			
	(0.001)				(0.000)			
Movie Released in Theaters	3.531***				-1.654***			
	(0.955)				(0.552)			
Number of Release Countries	0.044***				-0.022***			
	(0.015)				(0.006)			
Number of Awards Won	0.107***				-0.009***			
	(0.011)				(0.003)			
Elapsed Time since First Review			-0.866***	0.711***			0.342***	0.235**
			(0.114)	(0.249)			(0.052)	(0.117)
Review Position within Sequence			-0.051***	-0.052***			0.010***	0.010***
			(0.010)	(0.010)			(0.002)	(0.002)
Review Year				-0.575***				0.039
				(0.084)				(0.037)
Movie Fixed Effects	✗	✓	✓	✓	✗	✓	✓	✓
Reviewer Fixed Effects	✗	✗	✓	✓	✗	✗	✓	✓
R^2	0.068	0.182	0.457	0.457	0.018	0.065	0.262	0.262
Number of Observations	18578071	18578071	15825513	15825513	18578071	18578071	15825513	15825513

† A movie is considered female if the top actress is female. Specifications (1)-(4) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (5)-(8) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Box office revenue is measured in million USD. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 5: Comparison of Verified and Non-Verified Crowd Reviewers on Rotten Tomatoes.†

	Non-Verified Crowd Reviewers	Verified Crowd Reviewers	Difference	$p-$value
A. All Movies				
Extreme Low Rating Score	9.24	2.16	-7.09	0.000
Extreme Low VADER Score	1.91	0.30	-1.60	0.000
Extreme Low TextBlob Score	1.19	0.68	-0.51	0.000
Extreme Negative Tone	1.29	1.18	-0.11	0.000
B. Male Movies				
Extreme Low Rating Score	7.61	1.87	-5.74	0.000
Extreme Low VADER Score	1.51	0.26	-1.25	0.000
Extreme Low TextBlob Score	1.05	0.59	-0.45	0.000
Extreme Negative Tone	1.12	1.04	-0.09	0.002
C. Female Movies				
Extreme Low Rating Score	11.95	2.73	-9.22	0.000
Extreme Low VADER Score	2.57	0.39	-2.18	0.000
Extreme Low TextBlob Score	1.41	0.85	-0.56	0.000
Extreme Negative Tone	1.57	1.47	-0.10	0.023

† A movie is considered female if the top actress is female. The table reports the results of t-tests with equal variance between the groups of verified and non-verified Rotten Tomatoes crowd reviewers. The total number of non-verified and verified crowd reviews is equal to 407,806 and 465,645, respectively.

Table 6: Difference in Gender Score Gaps across Critics, Verified, and Non-Verified Rotten Tomatoes Crowds. †

Dependent variable:	Rating Score				Extreme Low Rating			
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	-2.607**	-2.984**			1.406***	1.572***		
	(1.278)	(1.437)			(0.470)	(0.574)		
Female Movie × Non Verified Crowd			-3.759***	-4.282***			2.135***	2.590***
			(1.302)	(1.456)			(0.604)	(0.788)
Female Movie × Verified Crowd			-0.432	-0.436			0.030	-0.388
			(1.645)	(1.828)			(0.439)	(0.689)
Non-Verified Crowd				3.036***				6.018***
				(1.066)				(0.448)
Verified Crowd				7.782***				4.218***
				(1.151)				(0.340)
Elapsed Time since First Review	-4.251***	-10.102***	-4.337***	-5.760***	2.453***	2.202*	2.508***	0.193
	(1.410)	(2.128)	(1.437)	(1.801)	(0.827)	(1.122)	(0.834)	(1.030)
Review Position within Sequence	-0.081***	-0.128***	-0.087***	-0.149***	-0.001	0.008	0.003	0.020***
	(0.030)	(0.047)	(0.029)	(0.036)	(0.010)	(0.005)	(0.009)	(0.006)
Review Year	-0.607	-0.746	-0.567	-0.521	0.047	-0.246	0.021	-0.362*
	(0.379)	(0.480)	(0.370)	(0.433)	(0.182)	(0.216)	(0.179)	(0.195)
Average Rating Assigned to Other Movies		0.392***		0.355***		-0.176***		-0.159***
		(0.022)		(0.019)		(0.015)		(0.013)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
R^2	0.567	0.264	0.567	0.274	0.459	0.086	0.459	0.089
Number of Observations	413490	459505	413490	459505	413490	459505	413490	459505

† A movie is considered female if the top actress is female. Specification (1)-(4) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (5)-(8) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

B Appendix

B.1 Institutional Details and Rescaling of Ratings

This section provides more details on the institutional setting of Rotten Tomatoes and IMDb, in particular to understand how each platform incorporates crowd ratings. Both Rotten Tomatoes provide a very simple way for users to leave ratings on their platform. On IMDb, a user is required to create an IMDb account in order to rate a movie, and they can report their gender within their personal account at any point. Note that a user's gender does not appear publicly on the platform, and that leaving a rating on IMDb does not require writing a review.[41] Posting a rating on Rotten Tomatoes works in a similar fashion, with some small differences. After creating an account, a user can easily leave a rating, with or without a review text. In either case, the rating will count towards the average audience score displayed on the movie's main page (under "Audience Score"). If the rating is posted without a review of at least 20 characters, however, the rating will not be displayed among the reviews listed under the "All Audience" page.[42] This means that the list of crowd ratings that I obtain from Rotten Tomatoes is a subset (i.e. the subset of ratings that were left with a review of at least 20 characters) of all the ratings left by users on Rotten Tomatoes. The list of ratings that appear in the IMDb crowd sample, however, consists of all the ratings left by IMDb users (without any review text).

Rating scores on IMDb and Rotten Tomatoes – coming from both critics and crowds – differ in their scales. Crowd ratings on IMDb are based on a discrete rating scale ranging from 1 to 10, while Rotten Tomatoes' crowd ratings range from 0.5 to 5. Given that Rotten Tomatoes' scale is simply half that of IMDb, I simply multiply Rotten Tomatoes ratings by 2 to make these two groups of ratings comparable. The ratings left by critics on Rotten Tomatoes differ more significantly, however. In particular, different critics rely on different rating scales that are sometimes non-numerical. To allow for a comparison of rating scores across critics and crowd reviewers on Rotten Tomatoes and IMDb, I proceed as follows to rescale all critics ratings to a common 1-10 scale. First, I use the conversion scale from Metacritic to convert non-numerical rating scales into a numerical 0-100 scale.[43] For each critic in the sample, I then convert their numerical ratings to a rating based on a 1-10 scale. More specifically, for each critic whose original scores S^O rely on a scale ranging from values S^O_{min} to S^O_{max}, I apply the following transformation to convert them into a scale ranging from 1 to 10:

$$S^{Scaled} = \frac{9 \times (S^O - S^O_{min})}{S^O_{max} - S^O_{min}} + 1$$

[41] IMDb separately reports user reviews as opposed to ratings. However the gender of the reviewer is not reported in this case, and some reviews only contain review text but no score. I therefore solely focus on the user ratings on IMDb.

[42] For more detailed information, see https://www.rottentomatoes.com/faq.

[43] See https://web.archive.org/web/20200626125515/https://www.metacritic.com/about-metascores.

For instance, for a critic whose scores range from $S^O_{min} = 1$ to $S^O_{max} = 5$, an original score of $S^O = 3$ would be converted to a score equal to $S^{Scaled} = \frac{9\times(3-1)}{5-1} + 1 = 5.5$.

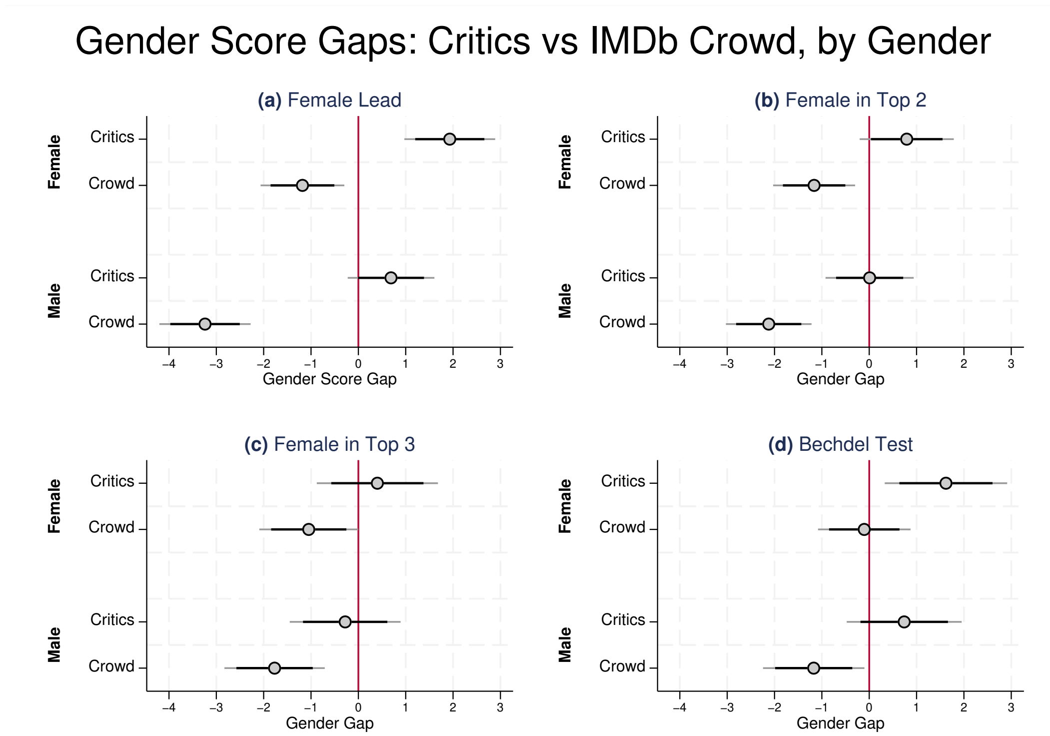
To test whether such rescaling has any impact on the main results of the analysis, I re-estimate the specifications presented in the main text for the subset of critics who rely on similar ratings scales as the crowd ratings. For instance, I restrict the sample to critics whose original rating scale ranges from 1 to 10, which implies that no rating transformation is needed to compare their ratings with those of the IMDb crowd. Performing such exercise results in quantitatively similar results despite relying on a very specific set of critics. Once the ratings from both critics and crowds on both platforms have been re-scaled to a common 1-10 range, I further scale them to a 10-100 range to facilitate the reading of the results in the main analysis.

B.2 Detailed Descriptives

As reported in Section 2.2 of the main text, a striking 79% of a movie's ratings comes from men among the IMDb crowd, on average. While this result is consistent with existing findings relying on similar data (Boyle, 2014; Stroube and Waguespack, 2024; Bayerl et al., 2024), it is interesting to see whether there exist differences in gender share of reviews across male and female movies. When focusing on movies with a male lead actor, the female share of reviews drops further to 18%, while the corresponding share for movies with a female lead actress is equal to 25.6%. These descriptives suggest that some gender-specific preferences might be at play, but identifying the mechanism driving these inequality in review posting is challenging. As reported by Bayerl et al. (2024), one potential explanation is that men and women might simply differ in their propensity to post online reviews. Another potential mechanism might relate to unequal time constraints between men and women. If women face more restrictions than men on their leisure time, their supply of movie consumption (and, by extension, of movie ratings) could be reduced. Charmes (2019) indeed reports that women participation in the labor market is constrained by a higher burden of participation in unpaid care work, which would naturally affect their participation in leisure-related activities. Given that an important determinant of time dedicated to unpaid care work is motherhood and care of children, one should perhaps observe differences across female reviewers of different age groups if leisure time constraints are driving the lower participation of women in movie-rating activity. The data indeed reveals that the female share of movies ratings declines further for reviewers of higher age groups. Among crowd reviewers less than 29 years old, the female share of reviews reaches 28%, while it declines to 20% and 17% among crowd reviewers aged between 30 and 44, and crowds over 45, respectively. Theses differences also appear when focusing on male and female movies separately. For female movies, the female share of crowd reviews reaches 35% among reviewers less than 29 and drops to 26% and 19% for reviewers aged 30-44 and over 45, respectively. For male movies, the corresponding female shares are 24%, 17%, and 15%. While these figures are consistent with stronger time-constraints put on women, they do not allow to rule out other motives for the gender disparities in movie reviewing.

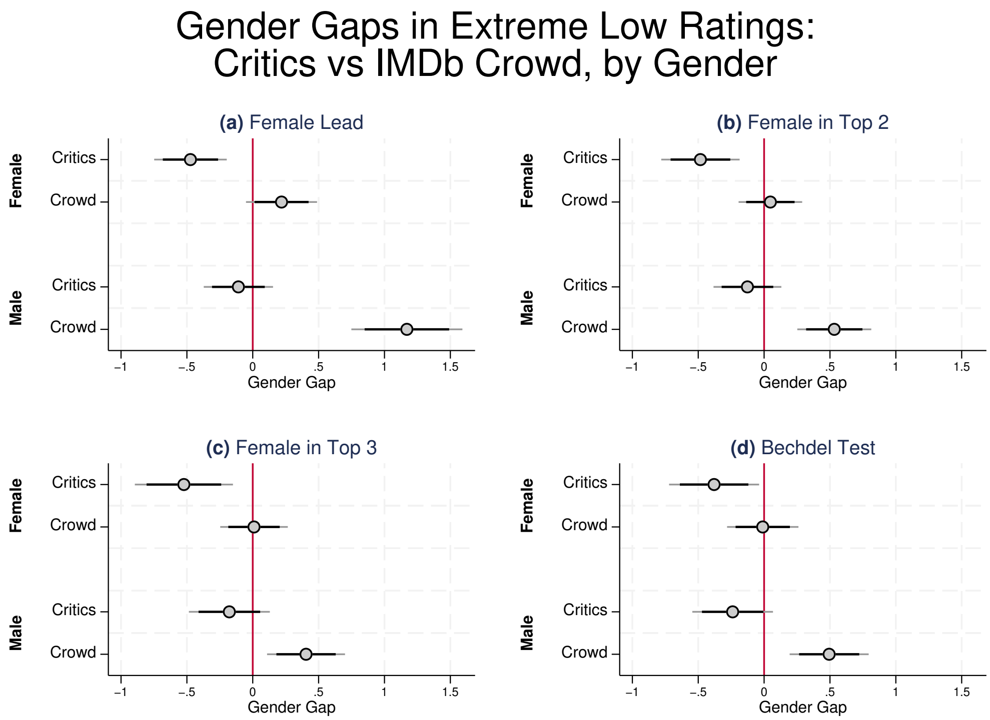
B.3 Estimation Results - Alternative Female Movie Definitions

The following reports estimation results based on IMDb and Rotten Tomatoes data relying on various female movie definitions as described in Section 3 in the main text.



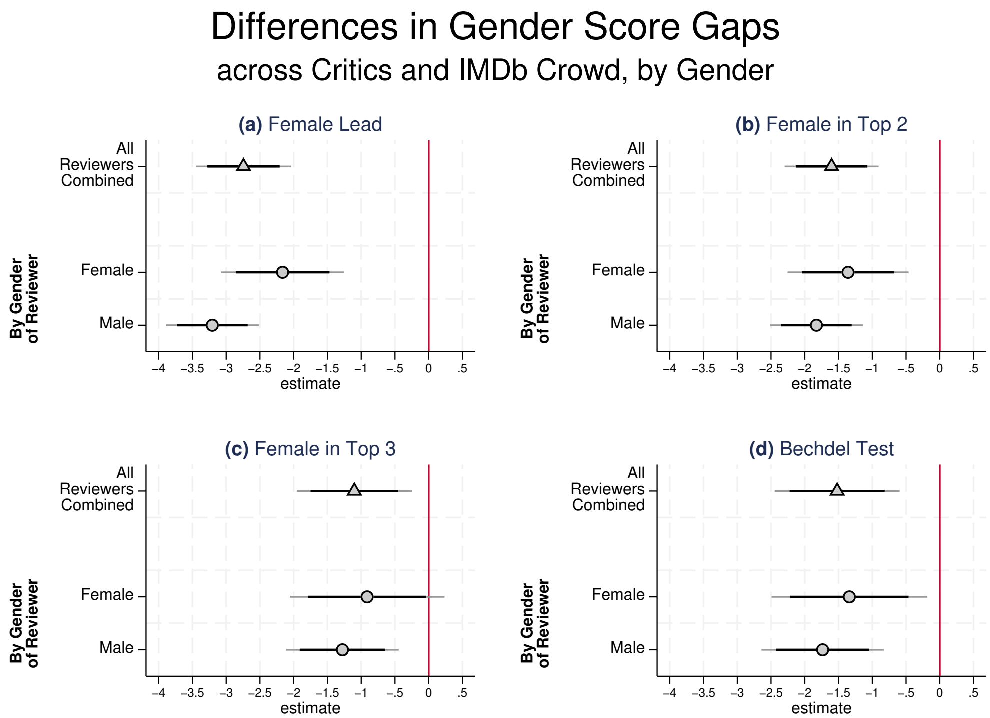
Note: For a given type of reviewer, each coefficient represents the movie gender score gap, which is defined as the difference in score given to female and male movies. Panel **(a)** considers a movie as female if the top lead actress is female. Panels **(b)** and **(c)** consider a movie as female if an actress is present in the Top 2 or Top 3 of the cast, respectively. Panel **(d)** considers a movie as female if it passes the Bechdel test. For each panel, all estimates come from a single regression as described in the text. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure A1: Gender Score Gap, by Reviewer Type and Gender.



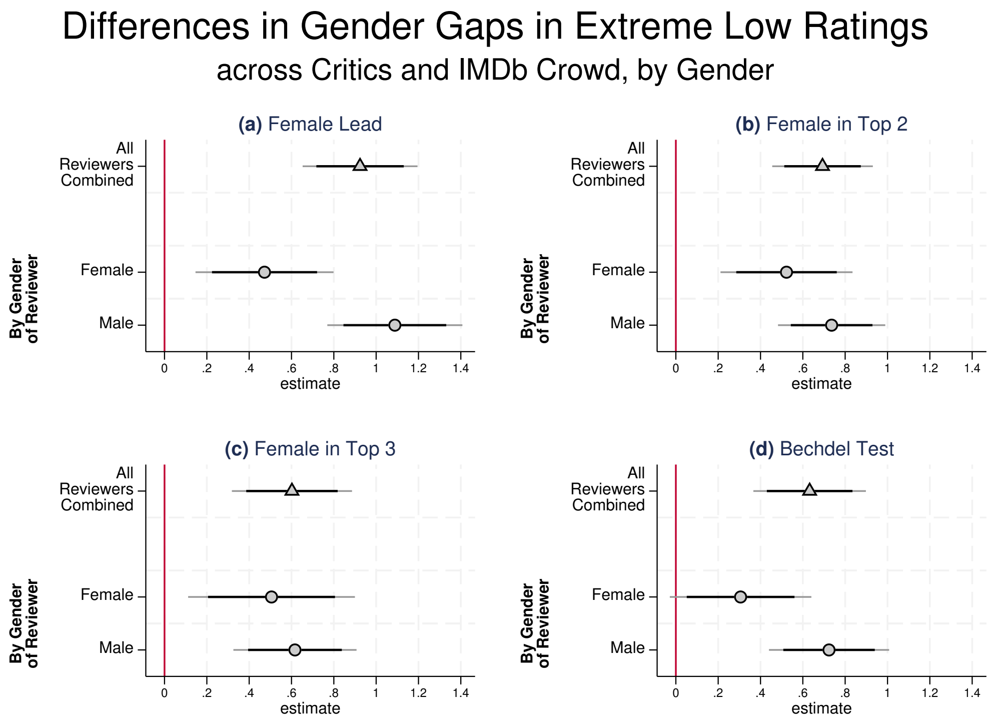
Note: For a given type of reviewer, each coefficient represents the movie gender gap in extreme low ratings, which is defined as the difference in the share of ratings scores of 10 given to female and male movies. Panel **(a)** considers a movie as female if the top lead actress is female. Panels **(b)** and **(c)** consider a movie as female if an actress is present in the Top 2 or Top 3 of the cast, respectively. Panel **(d)** considers a movie as female if it passes the Bechdel test. For each panel, all estimates come from a single regression as described in the text. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure A2: Gender Gaps in Extreme Low Ratings, by Reviewer Type and Gender.



Note: Each coefficient corresponds to a difference-in-differences estimate, which is defined as the difference in the gender score gaps across critics and audience reviewers. Panel **(a)** considers a movie as female if the top lead actress is female. Panels **(b)** and **(c)** consider a movie as female if an actress is present in the Top 2 or Top 3 of the cast, respectively. Panel **(d)** considers a movie as female if it passes the Bechdel test. For each type of reviewer, the gender score gap is itself defined as the difference in score given to female and male movies. For comparison, the difference-in-differences estimate for all reviewers combined is displayed in each panel. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure A3: Difference-in-Differences Score Gap Estimates, by Reviewers' Gender.



Note: Each coefficient corresponds to a difference-in-differences estimate, which is defined as the difference in the gender gap in extreme low ratings across critics and audience reviewers. Panel **(a)** considers a movie as female if the top lead actress is female. Panels **(b)** and **(c)** consider a movie as female if an actress is present in the Top 2 or Top 3 of the cast, respectively. Panel **(d)** considers a movie as female if it passes the Bechdel test. For each type of reviewer, the gender gap in extreme low ratings is itself defined as the difference in the share of rating scores of 10 given to female and male movies. For comparison, the difference-in-differences estimate for all reviewers combined is displayed in each panel. Standard errors are clustered at the movie level. Horizontal lines depict 99% and 95% confidence intervals.

Figure A4: Difference-in-Differences Extreme Low Ratings Estimates, by Reviewers' Gender.

Table A1: Gender Score Gaps between Critics and IMDb Crowds based on Various Female Movie Definitions. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie	0.986***		0.159		-0.154		0.919*	
	(0.352)		(0.360)		(0.455)		(0.470)	
Female Movie × Crowd Reviewer	-3.537***	-2.744***	-2.004***	-1.603***	-1.412***	-1.102***	-1.794***	-1.520***
	(0.367)	(0.273)	(0.353)	(0.270)	(0.427)	(0.331)	(0.436)	(0.359)
Crowd Reviewer	4.223***	2.362***	4.418***	2.499***	4.286***	2.353***	4.190***	2.456***
	(0.194)	(0.158)	(0.273)	(0.216)	(0.378)	(0.297)	(0.328)	(0.283)
Worldwide Box Office Revenue	0.005***		0.005***		0.005***		0.005***	
	(0.001)		(0.001)		(0.001)		(0.001)	
Movie Released in Theaters	1.688***		1.711***		1.753***		0.844	
	(0.565)		(0.570)		(0.573)		(0.999)	
Number of Release Countries	0.069***		0.070***		0.071***		0.038***	
	(0.010)		(0.010)		(0.010)		(0.012)	
Number of Awards Won	0.073***		0.072***		0.072***		0.071***	
	(0.008)		(0.008)		(0.008)		(0.008)	
Genre Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Production Studio Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
MPAA Rating Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Year of Release Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Origin Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Language Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Movie Fixed Effects	✗	✓	✗	✓	✗	✓	✗	✓
Gender Score Gap from Crowd	-2.550		-1.845		-1.566		-0.875	
p-value	0.000		0.000		0.000		0.028	
R^2	0.091	0.203	0.091	0.203	0.090	0.203	0.082	0.182
Number of Observations	337380397	337380397	337380397	337380397	337380397	337380397	283498730	283498730

† All specifications use the rating score (measured from 10 to 100) as the dependent variable. Box office revenue is measured in million USD. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A2: Gender Gaps in Extreme Low Ratings between Critics and IMDb Crowds based on Various Female Movie Definitions. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie	-0.213**		-0.207**		-0.254**		-0.276**	
	(0.097)		(0.097)		(0.117)		(0.115)	
Female Movie × Crowd Reviewer	1.124***	0.924***	0.641***	0.693***	0.584***	0.602***	0.664***	0.632***
	(0.137)	(0.105)	(0.106)	(0.092)	(0.119)	(0.110)	(0.110)	(0.103)
Crowd Reviewer	1.020***	1.546***	0.949***	1.399***	0.884***	1.360***	0.533***	0.887***
	(0.059)	(0.052)	(0.080)	(0.070)	(0.105)	(0.097)	(0.074)	(0.072)
Worldwide Box Office Revenue	0.000		0.000		0.000		0.000	
	(0.000)		(0.000)		(0.000)		(0.000)	
Movie Released in Theaters	-0.889***		-0.901***		-0.910***		-0.425	
	(0.316)		(0.317)		(0.317)		(0.431)	
Number of Release Countries	-0.031***		-0.032***		-0.032***		-0.018***	
	(0.003)		(0.003)		(0.003)		(0.003)	
Number of Awards Won	-0.004***		-0.004**		-0.004**		-0.004**	
	(0.002)		(0.002)		(0.002)		(0.002)	
Genre Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Production Studio Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
MPAA Rating Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Year of Release Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Origin Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Language Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
Movie Fixed Effects	✗	✓	✗	✓	✗	✓	✗	✓
Gender Gap from Crowd	0.912		0.433		0.331		0.388	
p-value	0.000		0.000		0.002		0.000	
R^2	0.011	0.071	0.011	0.071	0.011	0.071	0.007	0.040
Number of Observations	337380397	337380397	337380397	337380397	337380397	337380397	283498730	283498730

† All specifications use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Box office revenue is measured in million USD. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A3: Differences in Gender Gaps between Critics and Rotten Tomatoes Crowds based on Various Female Movie Definitions. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer	-1.785***	0.906***	-1.248***	0.802***	-0.821***	0.745***	-0.873***	0.393***
	(0.297)	(0.125)	(0.269)	(0.107)	(0.314)	(0.120)	(0.337)	(0.124)
Elapsed Time since First Review	0.711***	0.235**	0.707***	0.237**	0.709***	0.237**	0.831***	0.148
	(0.249)	(0.117)	(0.249)	(0.117)	(0.249)	(0.117)	(0.282)	(0.128)
Review Position within Sequence	-0.052***	0.010***	-0.052***	0.010***	-0.052***	0.010***	-0.052***	0.011***
	(0.010)	(0.002)	(0.010)	(0.002)	(0.010)	(0.002)	(0.010)	(0.002)
Review Year	-0.575***	0.039	-0.573***	0.038	-0.573***	0.038	-0.617***	0.053
	(0.084)	(0.037)	(0.084)	(0.037)	(0.084)	(0.037)	(0.094)	(0.041)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.457	0.262	0.457	0.262	0.457	0.262	0.449	0.265
Number of Observations	15825513	15825513	15825513	15825513	15825513	15825513	13418334	13418334

† For each Female movie definition, the first specification uses the rating score (measured from 10 to 100) as the dependent variable, while the second uses a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A4: Difference in Gender Score Gaps across Critics, Verified, and Non-Verified Rotten Tomatoes Crowds, based on Various Female Movie Definitions. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	-2.984**		-2.680		-2.299		0.806	
	(1.437)		(1.650)		(2.084)		(2.293)	
Female Movie × Non Verified Crowd		-4.282***		-4.209**		-2.472		0.178
		(1.456)		(1.697)		(2.037)		(2.196)
Female Movie × Verified Crowd		-0.436		0.371		-1.824		1.532
		(1.828)		(1.996)		(2.351)		(2.734)
Non-Verified Crowd		3.036***		4.200***		3.563*		0.592
		(1.066)		(1.532)		(1.878)		(1.905)
Verified Crowd		7.782***		7.526***		9.090***		5.566**
		(1.151)		(1.639)		(2.117)		(2.426)
Elapsed Time since First Review	-10.102***	-5.760***	-10.165***	-6.122***	-10.160***	-5.806***	-10.150***	-5.607***
	(2.128)	(1.801)	(2.127)	(1.806)	(2.126)	(1.790)	(2.469)	(2.063)
Review Position within Sequence	-0.128***	-0.149***	-0.129***	-0.149***	-0.128***	-0.142***	-0.125***	-0.139***
	(0.047)	(0.036)	(0.047)	(0.038)	(0.046)	(0.038)	(0.045)	(0.037)
Review Year	-0.746	-0.521	-0.724	-0.473	-0.736	-0.539	-0.700	-0.560
	(0.480)	(0.433)	(0.481)	(0.434)	(0.482)	(0.448)	(0.565)	(0.520)
Average Rating Assigned to Other Movies	0.392***	0.355***	0.392***	0.355***	0.392***	0.355***	0.383***	0.347***
	(0.022)	(0.019)	(0.022)	(0.019)	(0.022)	(0.019)	(0.025)	(0.022)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✗	✗	✗	✗	✗	✗	✗	✗
R^2	0.264	0.274	0.264	0.274	0.264	0.273	0.252	0.261
Number of Observations	459505	459505	459505	459505	459505	459505	377284	377284

† All specifications use the rating score (measured from 10 to 100) as the dependent variable. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A5: Difference in Gender Gaps in Extreme Low Ratings across Critics, Verified, and Non-Verified Rotten Tomatoes Crowds, based on Various Female Movie Definitions. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	1.572***		1.555***		1.526*		0.718	
	(0.574)		(0.568)		(0.858)		(0.716)	
Female Movie × Non Verified Crowd		2.590***		2.499***		2.192**		1.576*
		(0.788)		(0.742)		(1.035)		(0.885)
Female Movie × Verified Crowd		-0.388		-0.274		0.276		-0.721
		(0.689)		(0.590)		(0.761)		(0.684)
Non-Verified Crowd		6.018***		5.342***		5.098***		5.673***
		(0.448)		(0.592)		(0.958)		(0.635)
Verified Crowd		4.218***		4.178***		3.894***		4.530***
		(0.340)		(0.431)		(0.708)		(0.577)
Elapsed Time since First Review	2.202*	0.193	2.235**	0.400	2.231**	0.281	2.508*	0.386
	(1.122)	(1.030)	(1.119)	(1.034)	(1.116)	(1.018)	(1.298)	(1.175)
Review Position within Sequence	0.008	0.020***	0.008*	0.019***	0.008	0.015***	0.005	0.013**
	(0.005)	(0.006)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Review Year	-0.246	-0.362*	-0.258	-0.383*	-0.251	-0.354*	-0.160	-0.230
	(0.216)	(0.195)	(0.215)	(0.196)	(0.216)	(0.201)	(0.246)	(0.224)
Average Rating Assigned to Other Movies	-0.176***	-0.159***	-0.176***	-0.159***	-0.177***	-0.159***	-0.175***	-0.158***
	(0.015)	(0.013)	(0.015)	(0.013)	(0.015)	(0.013)	(0.018)	(0.016)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✗	✗	✗	✗	✗	✗	✗	✗
R^2	0.086	0.089	0.086	0.089	0.086	0.089	0.079	0.083
Number of Observations	459505	459505	459505	459505	459505	459505	377284	377284

† All specifications use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A6: Difference in Gender Score Gaps across Critics, Verified, and Non-Verified Rotten Tomatoes Crowd. VADER Sentiment Analysis. †

Dependent Variable:	Rating Score				Extreme Low Rating			
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	-0.029*	-0.028*			0.361**	0.422***		
	(0.015)	(0.015)			(0.176)	(0.156)		
Female Movie × Non Verified Crowd			-0.045***	-0.046***			0.620***	0.716***
			(0.017)	(0.017)			(0.233)	(0.211)
Female Movie × Verified Crowd			0.002	0.007			-0.124	-0.139
			(0.019)	(0.020)			(0.161)	(0.176)
Non-Verified Crowd				0.181***				1.353***
				(0.010)				(0.091)
Verified Crowd				0.193***				0.713***
				(0.011)				(0.071)
Elapsed Time since First Review	-0.102***	-0.110***	-0.103***	-0.088***	1.235***	1.785***	1.253***	1.118***
	(0.023)	(0.023)	(0.023)	(0.021)	(0.423)	(0.372)	(0.425)	(0.353)
Review Position within Sequence	-0.001***	-0.002***	-0.001***	-0.002***	0.003	0.006***	0.004	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.002)	(0.004)	(0.004)
Review Year	-0.002	-0.001	-0.001	0.000	-0.190*	0.030	-0.199*	-0.007
	(0.006)	(0.006)	(0.006)	(0.006)	(0.104)	(0.085)	(0.105)	(0.083)
Average Rating Assigned to Other Movies		0.004***		0.004***		-0.027***		-0.021***
		(0.000)		(0.000)		(0.003)		(0.002)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
R^2	0.404	0.127	0.404	0.128	0.274	0.015	0.274	0.017
Number of Observations	396562	444329	396562	444329	396562	444329	396562	444329

† A movie is considered female if the top actress is female. All specifications rely on the VADER classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd reviews come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A7: Difference in Gender Score Gaps across Critics, Verified, and Non-Verified Rotten Tomatoes Crowd. TextBlob Sentiment Analysis. †

Dependent Variable:	Rating Score				Extreme Low Rating			
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	-0.018**	-0.020**			0.170	0.169		
	(0.008)	(0.009)			(0.123)	(0.113)		
Female Movie × Non Verified Crowd			-0.025***	-0.029***			0.222*	0.229*
			(0.009)	(0.009)			(0.133)	(0.124)
Female Movie × Verified Crowd			-0.006	-0.002			0.071	0.057
			(0.011)	(0.013)			(0.126)	(0.122)
Non-Verified Crowd				0.063***				0.552***
				(0.007)				(0.078)
Verified Crowd				0.117***				0.561***
				(0.008)				(0.079)
Elapsed Time since First Review	-0.011	-0.094***	-0.012	-0.048***	0.607*	0.352	0.610*	0.317
	(0.011)	(0.015)	(0.011)	(0.012)	(0.354)	(0.284)	(0.353)	(0.274)
Review Position within Sequence	-0.000	-0.000	-0.000	-0.001***	0.005*	0.006***	0.005**	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.002)	(0.002)	(0.001)
Review Year	-0.005**	-0.006	-0.005*	-0.004	-0.020	-0.149**	-0.022	-0.152**
	(0.003)	(0.004)	(0.003)	(0.003)	(0.078)	(0.063)	(0.078)	(0.063)
Average Rating Assigned to Other Movies		0.003***		0.002***		-0.018***		-0.018***
		(0.000)		(0.000)		(0.002)		(0.002)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
R^2	0.436	0.134	0.436	0.142	0.309	0.009	0.309	0.009
Number of Observations	396562	444329	396562	444329	396562	444329	396562	444329

† A movie is considered female if the top actress is female. All specifications rely on the TextBlob classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd reviews come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A8: Difference in Gender Score Gaps across Critics, Verified, and Non-Verified Rotten Tomatoes Crowd. LIWC Sentiment Analysis. †

Dependent Variable:	Rating Score				Extreme Low Rating			
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd	0.054	0.079			0.260**	0.266***		
	(0.130)	(0.126)			(0.111)	(0.098)		
Female Movie × Non Verified Crowd			0.152	0.179			0.306***	0.322***
			(0.137)	(0.135)			(0.118)	(0.105)
Female Movie × Verified Crowd			-0.129	-0.110			0.174	0.168
			(0.141)	(0.144)			(0.125)	(0.130)
Non-Verified Crowd				0.309***				0.707***
				(0.091)				(0.063)
Verified Crowd				0.138*				0.995***
				(0.083)				(0.078)
Elapsed Time since First Review	0.235	0.195	0.241	0.002	0.167	-0.616***	0.170	-0.445*
	(0.178)	(0.169)	(0.178)	(0.152)	(0.286)	(0.235)	(0.286)	(0.234)
Review Position within Sequence	0.004	0.005**	0.005	0.006**	0.004	0.005*	0.004	0.005*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Review Year	0.091**	0.084*	0.088*	0.073	0.110	0.033	0.108	0.039
	(0.046)	(0.048)	(0.045)	(0.047)	(0.073)	(0.068)	(0.073)	(0.069)
Average Rating Assigned to Other Movies		-0.023***		-0.022***		-0.016***		-0.017***
		(0.002)		(0.002)		(0.002)		(0.002)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✗	✓	✗	✓	✗	✓	✗
R^2	0.361	0.065	0.361	0.066	0.321	0.008	0.321	0.008
Number of Observations	396501	444267	396501	444267	396501	444267	396501	444267

† A movie is considered female if the top actress is female. All specifications rely on the LIWC's negative tone dictionary to construct reviews' sentiment score, ranging from 0 to 100. A higher score reflects a more negative review sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 99^{th} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd reviews come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

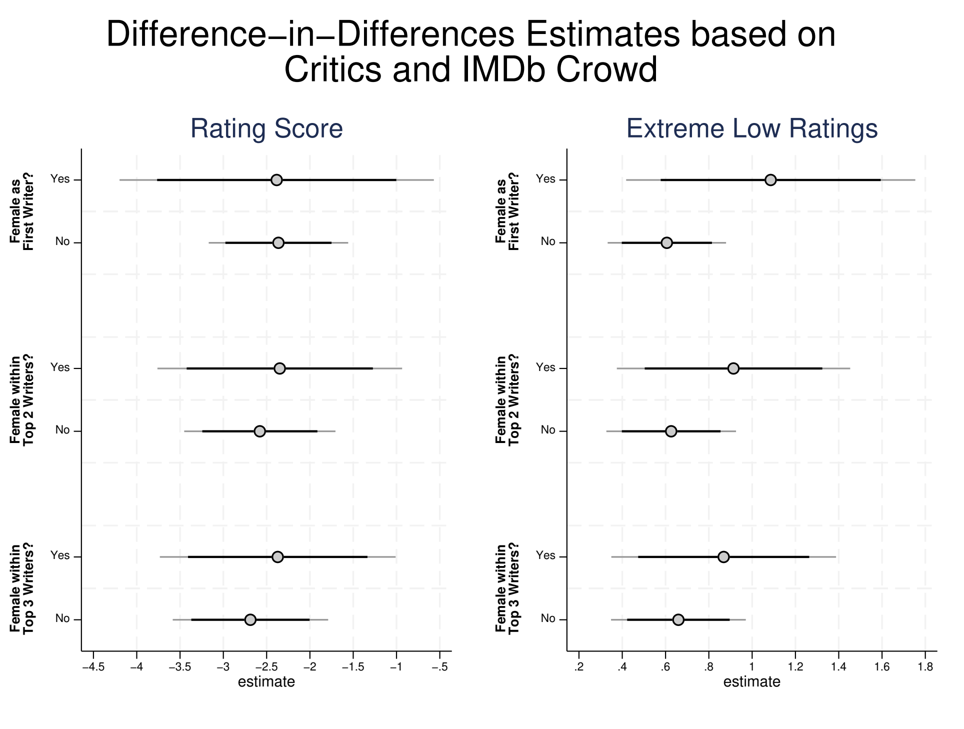
B.4 Unobservable Differences in Movie Content

This section reports the results of performing similar estimations as in Sections 4.1 and 4.2 of the main text, but within different subsets of movies. More specifically, I consider the subset of female-written movies and male-written movies as well as the subset of male-produced and female-produced movies separately.

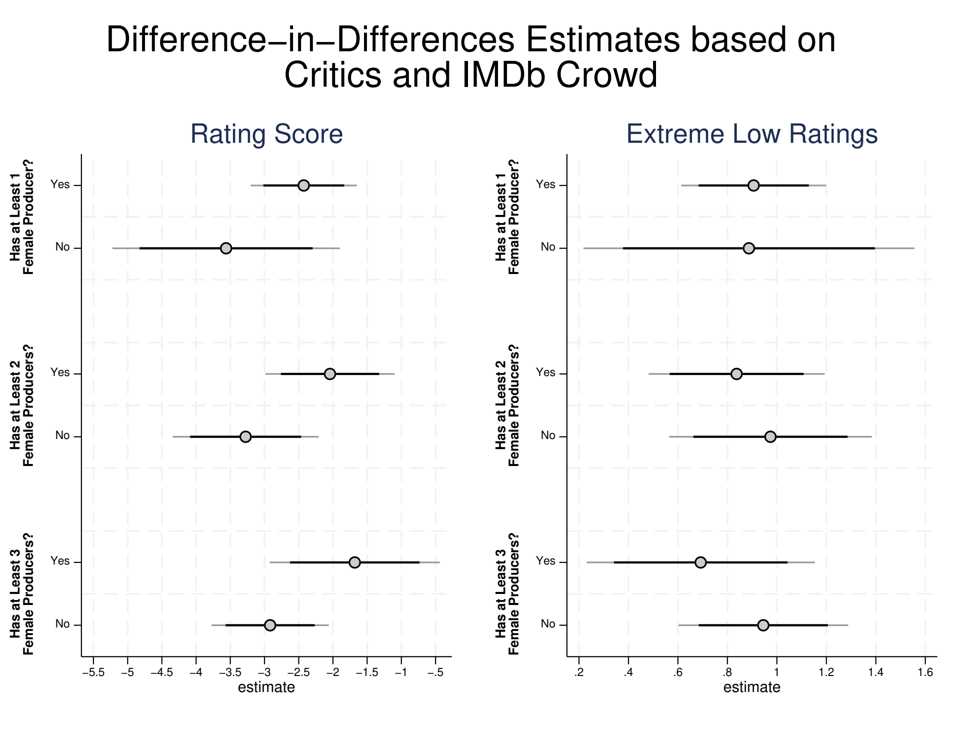
Relying on the IMDb data, I start by estimating equation (1) within the subset of female-written movies and male-written movies separately. In other words, I ask whether the increase in gender gaps between critics and crowds is found among both male-written movies and female-written movies. I perform a similar exercise comparing male produced and female produced movies. IMDb provides the list of each movie's producers as well as screen writers. Screen writers are listed in an order that reflects level of contribution.[44] I can therefore distinguish between movies whose first writer is female, movies that have a female writer among their top 2 writers, and movies that have a female writer among their top 3 writers. For movie producers, IMDb lists them in alphabetical order and does not provide information on their individual level of contribution. I therefore distinguish between movies that have at least 3, at least 2, or at least 1 female producers among their producer team. Figure A5a reports the difference-in-differences coefficients distinguishing between male and female written movies. For instance, the top estimates reported in both graphs rely on the subset of movies with a female top writer (i.e. "female-written movies"), while the second estimates focus on the subset of movies with a male top writer (i.e. "male-written movies"). The remainder of estimates focus on samples constructed with alternative definitions of a female-written movies (based on whether a female is among the top 2 or the top 3 writers). The figure indicates that the difference-in-differences estimates are similar across all subsamples, providing support to the fact that the increase in the gender gaps documented above are driven by differences in female presence among the top cast rather than differences in content. As an alternative test, Figure A5b reports similar estimates focusing on subset of movies based on the gender of their producers. Again, the results indicate that the increase in the gender gaps are present both within the sets of male and female produced movies. Tables A9 and A10 report the complete results of all regressions.

I perform a similar exercise relying on the Rotten Tomatoes data. Figure A6a reports the corresponding difference-in-differences estimates when considering female and male-written movie separately. Figure A6b performs the same exercise considering male and female-produced movies. In all cases, the resulting estimates are very similar across all subsamples. These results again provide support to the fact that the increase in gender gaps are driven by differences in movies' on-screen female presence rather than differences in movie content. Tables A11 and A12 report the complete estimation results of all the corresponding regressions.

[44] See https://help.imdb.com/article/contribution/filmography-credits/writers/GPLAT3NTCGA67A6R?ref_=helpms_helpart_inline#.

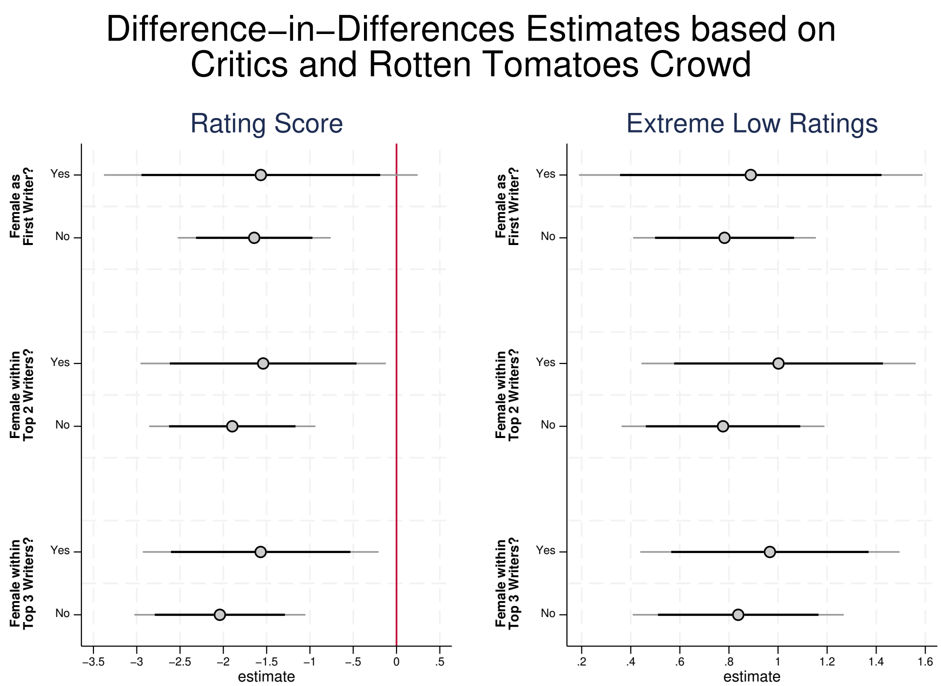


(a) By Gender of Writer.

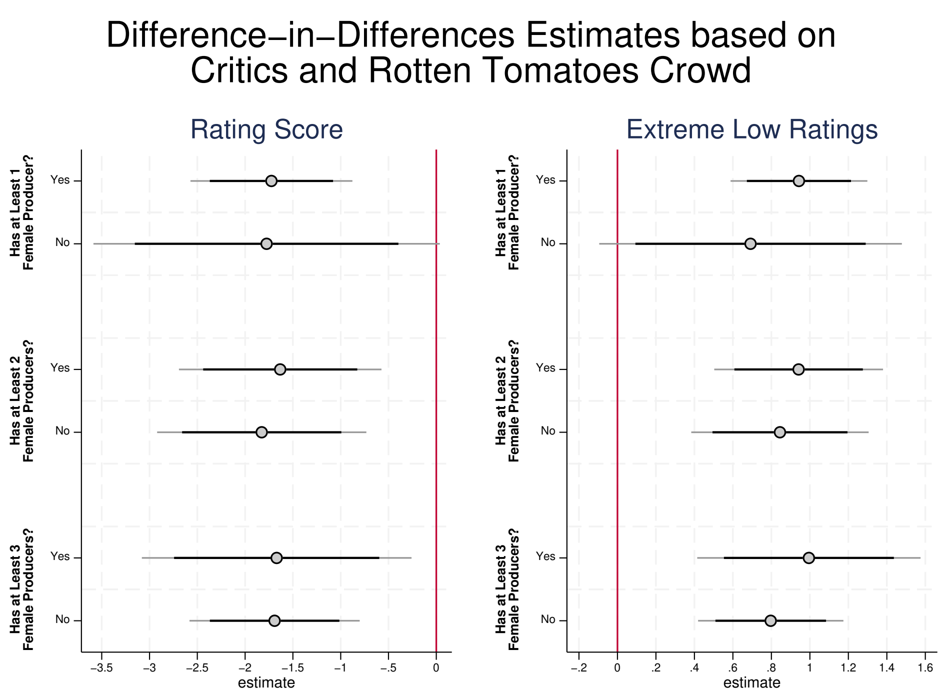


(b) By Gender of Producer.

Figure A5: Difference-in-Differences Estimates based on Critics and IMDb Crowds, by Gender of and Writer and Producer.



(a) By Gender of Writer.



(b) By Gender of Producer.

Figure A6: Difference-in-Differences Estimates based on Critics and Rotten Tomatoes Crowds, by Gender of Writer and Producer.

Table A9: Gender Gaps between Critics and IMDb Crowds, by Gender of Writer. †

Dependent Variable:	Rating Score			Extreme Low Rating		
Female-Written Movie Based on:	Female as Top Writer	Female among Top 2 Writers	Female among Top 3 Writers	Female as Top Writer	Female among Top 2 Writers	Female among Top 3 Writers
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
Crowd	2.495***	2.506***	2.480***	1.490***	1.458***	1.454***
	(0.165)	(0.172)	(0.174)	(0.055)	(0.057)	(0.058)
Female Movie × Crowd	-2.363***	-2.578***	-2.687***	0.606***	0.626***	0.659***
	(0.313)	(0.339)	(0.348)	(0.106)	(0.116)	(0.121)
Female Movie × Crowd × Female-Written	-0.020	0.230	0.315	0.480*	0.288	0.209
	(0.771)	(0.645)	(0.633)	(0.280)	(0.239)	(0.235)
Crowd × Female-Written	-1.413**	-0.906**	-0.712	0.634***	0.548***	0.510***
	(0.592)	(0.452)	(0.440)	(0.165)	(0.142)	(0.140)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓
Diff-in-Diff for Female-Written Movies	-2.384	-2.348	-2.372	1.086	0.914	0.869
p-value	0.001	0.000	0.000	0.000	0.000	0.000
R^2	0.203	0.203	0.204	0.071	0.071	0.071
Number of Observations	335626056	333678781	332543004	335626056	333678781	332543004

† A movie is considered female if the top actress is female. Specifications (1)-(3) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (4)-(6) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A10: Gender Gaps between Critics and IMDb Crowds, by Gender of Producer. †

Dependent Variable:	Rating Score			Extreme Low Rating		
Female-Produced Movie Based on:	# Female Producers ≥ 1	# Female Producers ≥ 2	# Female Producers ≥ 3	# Female Producers ≥ 1	# Female Producers ≥ 2	# Female Producers ≥ 3
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
Crowd	3.924***	3.200***	2.864***	1.453***	1.461***	1.439***
	(0.340)	(0.219)	(0.182)	(0.131)	(0.076)	(0.058)
Female Movie × Crowd	-3.561***	-3.276***	-2.917***	0.887***	0.974***	0.945***
	(0.645)	(0.414)	(0.333)	(0.260)	(0.159)	(0.133)
Female Movie × Crowd × Female-Produced	1.135	1.233**	1.237**	0.019	-0.137	-0.253
	(0.712)	(0.553)	(0.586)	(0.283)	(0.211)	(0.223)
Crowd × Female-Produced	-1.990***	-1.708***	-2.014***	0.117	0.172*	0.430***
	(0.383)	(0.314)	(0.361)	(0.142)	(0.104)	(0.127)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓
Diff-in-Diff for Female-Produced Movies	-2.426	-2.043	-1.681	0.906	0.837	0.692
p-value	0.000	0.000	0.000	0.000	0.000	0.000
R^2	0.203	0.203	0.203	0.071	0.071	0.071
Number of Observations	337155930	337155930	337155930	337155930	337155930	337155930

† A movie is considered female if the top actress is female. Specifications (1)-(3) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (4)-(6) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A11: Gender Gaps between Critics and Rotten Tomatoes Crowds, by Gender of Writer. †

Dependent Variable:	Rating Score			Extreme Low Rating		
Female-Written Movie Based on:	Female as Top Writer	Female among Top 2 Writers	Female among Top 3 Writers	Female as Top Writer	Female among Top 2 Writers	Female among Top 3 Writers
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
Female Movie × Crowd	-1.643***	-1.897***	-2.040***	0.782***	0.776***	0.838***
	(0.343)	(0.373)	(0.383)	(0.144)	(0.160)	(0.167)
Female Movie × Crowd × Female-Written	0.076	0.357	0.471	0.107	0.226	0.129
	(0.781)	(0.662)	(0.649)	(0.307)	(0.269)	(0.262)
Elapsed Time since First Review	0.712***	0.703***	0.709***	0.235**	0.238**	0.237**
	(0.250)	(0.251)	(0.250)	(0.117)	(0.118)	(0.117)
Review Position within Sequence	-0.052***	-0.052***	-0.052***	0.010***	0.010***	0.010***
	(0.010)	(0.010)	(0.010)	(0.002)	(0.002)	(0.002)
Review Year	-0.577***	-0.573***	-0.575***	0.040	0.039	0.039
	(0.084)	(0.085)	(0.084)	(0.037)	(0.037)	(0.037)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓
Diff-in-Diff for Female-Written Movies	-1.567	-1.540	-1.569	0.889	1.002	0.967
p-value	0.026	0.005	0.003	0.001	0.000	0.000
R^2	0.457	0.457	0.457	0.262	0.262	0.262
Number of Observations	15757665	15694821	15663870	15757665	15694821	15663870

† A movie is considered female if the top actress is female. Specifications (1)-(3) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (4)-(6) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A12: Gender Gaps between Critics and Rotten Tomatoes Crowds, by Gender of Producer. †

Dependent Variable:	Rating Score			Extreme Low Rating		
Female-Produced Movie Based on:	# Female Producers ≥ 1	# Female Producers ≥ 2	# Female Producers ≥ 3	# Female Producers ≥ 1	# Female Producers ≥ 2	# Female Producers ≥ 3
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
Female Movie × Crowd	-1.775**	-1.826***	-1.691***	0.692**	0.845***	0.797***
	(0.703)	(0.425)	(0.346)	(0.305)	(0.179)	(0.147)
Female Movie × Crowd × Female-Produced	0.050	0.193	0.022	0.251	0.097	0.198
	(0.777)	(0.589)	(0.643)	(0.336)	(0.245)	(0.266)
Elapsed Time since First Review	0.705***	0.704***	0.704***	0.236**	0.236**	0.238**
	(0.249)	(0.249)	(0.249)	(0.117)	(0.117)	(0.117)
Review Position within Sequence	-0.052***	-0.052***	-0.052***	0.010***	0.010***	0.010***
	(0.010)	(0.010)	(0.010)	(0.002)	(0.002)	(0.002)
Review Year	-0.574***	-0.574***	-0.574***	0.039	0.039	0.039
	(0.084)	(0.084)	(0.084)	(0.037)	(0.037)	(0.037)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓
Diff-in-Diff for Female-Produced Movies	-1.725	-1.633	-1.670	0.943	0.942	0.995
p-value	0.000	0.000	0.002	0.000	0.000	0.000
R^2	0.457	0.457	0.457	0.262	0.262	0.262
Number of Observations	15822004	15822004	15822004	15822004	15822004	15822004

† A movie is considered female if the top actress is female. Specifications (1)-(3) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (4)-(6) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Crowd ratings come from IMDb. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

B.5 Sentiment Analysis

I rely on three rule-based analyzers to construct measures of each review's sentiment. I first rely on VADER (Valence Aware Dictionary for sEntiment Reasoning) and TextBlob's sentiment analyzer, two popular tools that provide measures of the sentiment of a given text, ranging from -1 (most negative sentiment) to 1 (more positive sentiment). I additionally use the LIWC (Linguistic Inquiry and Word Count) tool to construct additional measures of text sentiment (Pennebaker et al., 2001; Boyd et al., 2022).[45] For a given text, LIWC classifies words in various pre-defined categories (based on linguistic dimensions, psychological processes, etc.) including positive and negative emotions. I use the LIWC's negative tone dictionary to obtain a measure of the percentage of negative-tone words (e.g. "bad," "hate," "wrong" etc.) out of a given review. This measures accordingly ranges from 0 (no words with negative tone) to 100 (all words with negative tone).

Specifications (1)-(3) of Table A13 reports the results of estimating equation (3) using the three sentiment score measures (VADER, TextBlob, and LIWC) as the dependent variable. Each specification includes both movie and reviewer fixed effects and additionally controls for the sequential and temporal dynamics of ratings. The estimates show very similar results as the ones relying on rating scores. Based on the three sentiment score measures, the difference-in-differences estimates indicate that the gender sentiment score gap is significantly larger among crowd reviewers than among critics, and at the expense of female movies.

In order to test whether extreme behavior is also reflected within review's sentiment scores, I further construct measures of extreme low sentiment scores. For the VADER and TextBlob classifiers, I define sentiment scores as extremely low if they fall in the 1^{st} percentile of the respective review sentiment score distribution. For the LIWC classifier, I define sentiment scores as extremely low if they fall in the 99^{th} percentile of the review sentiment score distribution.[46] Specifications (4)-(6) present the results of using these measures as dependent variables. Consistent with the results presented in Table 4, the estimates show that crowds increase the gender gap in extreme low sentiment scores relative to critics. Overall, the estimates presented in Table A13 provide robustness to our main results by showing that crowd reviewers increase gender gaps in movie reviews' sentiment relative to critics.[47]

[45] While LIWC is a very popular text analysis tool, it is well-established that VADER performs better in the context of social media and online text. In particular, "VADER distinguishes itself from LIWC in that it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domains" (Hutto and Gilbert, 2014).

[46] Recall that higher values of the LIWC sentiment score reflect more negative reviews.

[47] Tables A14, A15, and A16 further provide the results of estimating similar specifications using alternative definitions of female movies based on the VADER, TextBlob, and LIWC sentiment scores, respectively.

Table A13: Difference in Gender Sentiment Gaps across Critics and Rotten Tomatoes Crowds. †

	Sentiment Score			Extreme Sentiment Score		
Sentiment Classifier:	VADER	TextBlob	LIWC	VADER	TextBlob	LIWC
	(1)	(2)	(3)	(4)	(5)	(6)
	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.
Female Movie × Crowd Reviewer	-0.028***	-0.016***	0.136***	0.223***	0.126***	0.142***
	(0.004)	(0.002)	(0.030)	(0.056)	(0.035)	(0.035)
Elapsed Time since First Review	-0.009**	-0.003	0.107***	0.103**	0.177***	0.217***
	(0.004)	(0.002)	(0.028)	(0.043)	(0.041)	(0.043)
Review Position within Sequence	-0.001***	-0.000***	0.003***	0.001	0.004***	0.004***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Review Year	-0.005***	-0.003***	0.008	0.014	-0.013	-0.011
	(0.001)	(0.001)	(0.009)	(0.014)	(0.014)	(0.015)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓
R^2	0.262	0.281	0.213	0.138	0.165	0.165
Number of Observations	14703460	14703460	14701429	14703460	14703460	14701429

† A movie is considered female if the top actress is female. Specifications (1)-(3) use the review sentiment score from the corresponding classifier as the dependent variable. Both the VADER and TextBlob score range from -1 to 1, and a larger score reflects a more positive sentiment. Specifications (3) and (6) rely on LIWC's negative tone dictionary to construct reviews' sentiment score, ranging from 0 to 100. A higher score reflects a more negative review sentiment. Specifications (4)-(6) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. For the VADER and TextBlob classifiers, extreme scores are defined as scores that fall in the 1^{st} percentile of the respective review sentiment score distribution. For the LIWC classifier, extreme scores are defined as scores that fall in the 99^{th} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd scores come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A14: Differences in Gender Gaps between Critics and Rotten Tomatoes Crowd based on Various Female Movie Definitions. VADER Sentiment Analysis. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer	-0.028***	0.223***	-0.023***	0.150***	-0.027***	0.098*	-0.016***	0.096*
	(0.004)	(0.056)	(0.004)	(0.048)	(0.004)	(0.059)	(0.004)	(0.057)
Elapsed Time since First Review	-0.009**	0.103**	-0.009**	0.103**	-0.009**	0.103**	-0.007	0.093**
	(0.004)	(0.043)	(0.004)	(0.043)	(0.004)	(0.043)	(0.005)	(0.046)
Review Position within Sequence	-0.001***	0.001	-0.001***	0.001	-0.001***	0.001	-0.001***	0.001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Review Year	-0.005***	0.014	-0.005***	0.013	-0.005***	0.013	-0.006***	0.016
	(0.001)	(0.014)	(0.001)	(0.014)	(0.001)	(0.014)	(0.001)	(0.015)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.262	0.138	0.262	0.138	0.262	0.138	0.266	0.140
Number of Observations	14703460	14703460	14703460	14703460	14703460	14703460	12469125	12469125

† All specifications rely on the VADER classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. For each Female movie definition, the first specification uses the review sentiment score as the dependent variable, while the second uses a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A15: Differences in Gender Gaps between Critics and Rotten Tomatoes Crowd based on Various Female Movie Definitions. TextBlob Sentiment Analysis. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer	-0.016***	0.126***	-0.015***	0.133***	-0.014***	0.104***	-0.008***	0.039
	(0.002)	(0.035)	(0.002)	(0.033)	(0.002)	(0.040)	(0.002)	(0.039)
Elapsed Time since First Review	-0.003	0.177***	-0.003	0.177***	-0.003	0.177***	-0.001	0.144***
	(0.002)	(0.041)	(0.002)	(0.041)	(0.002)	(0.041)	(0.002)	(0.045)
Review Position within Sequence	-0.000***	0.004***	-0.000***	0.004***	-0.000***	0.004***	-0.000***	0.004***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Review Year	-0.003***	-0.013	-0.003***	-0.014	-0.003***	-0.014	-0.004***	-0.004
	(0.001)	(0.014)	(0.001)	(0.014)	(0.001)	(0.014)	(0.001)	(0.015)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.281	0.165	0.281	0.165	0.281	0.165	0.285	0.176
Number of Observations	14703460	14703460	14703460	14703460	14703460	14703460	12469125	12469125

† All specifications rely on the TextBlob classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. For each Female movie definition, the first specification uses the review sentiment score as the dependent variable, while the second uses a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A16: Differences in Gender Gaps between Critics and Rotten Tomatoes Crowd based on Various Female Movie Definitions. LIWC Sentiment Analysis. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score	Sentiment Score	Extreme Sentiment Score
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer	0.136***	0.142***	0.142***	0.148***	0.164***	0.142***	0.093***	0.076**
	(0.030)	(0.035)	(0.029)	(0.033)	(0.034)	(0.039)	(0.033)	(0.038)
Elapsed Time since First Review	0.107***	0.217***	0.107***	0.218***	0.107***	0.218***	0.079**	0.197***
	(0.028)	(0.043)	(0.028)	(0.043)	(0.028)	(0.043)	(0.032)	(0.048)
Review Position within Sequence	0.003***	0.004***	0.003***	0.004***	0.003***	0.004***	0.003***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Review Year	0.008	-0.011	0.008	-0.012	0.008	-0.012	0.015	-0.004
	(0.009)	(0.015)	(0.009)	(0.015)	(0.009)	(0.015)	(0.010)	(0.016)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.213	0.165	0.213	0.165	0.213	0.165	0.219	0.175
Number of Observations	14701429	14701429	14701429	14701429	14701429	14701429	12467392	12467392

† All specifications rely on the LIWC's negative tone dictionary to construct reviews' sentiment score, ranging from 0 to 100. A higher score reflects a more negative review sentiment. For each Female movie definition, the first specification uses the review sentiment score as the dependent variable, while the second uses a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 99^{th} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

B.6 Dynamics

To test whether reviewers who post a rating are reacting to the prevailing review environment, I follow Moe and Trusov (2011) and Godes and Silva (2012) and start by including two additional variables in equation (3): the mean and the standard deviation of all previous ratings of a given movie. The estimation results are presented in Table A17. In line with Godes and Silva (2012) – and consistent with herding behavior among reviewers – the estimates show that higher prevailing mean ratings are associated with higher subsequent ratings. On the other hand, more noisy prevailing ratings (i.e. a higher standard deviation of prevailing ratings) are associated with lower subsequent ratings. Note that the difference-in-differences estimates remain significant even controlling for these additional variables.[48]

To further account for the possibility of herding behavior among reviewers, I now explore whether the differences in the gender gaps across critics and crowds vary along a movie's life (Godes and Mayzlin, 2004). I do so by distinguishing between reviews posted in the early and late period of a movie's life and by relying on various cutoffs to define these two periods. More specifically, I define the following dummy variable:

$$Early_{it} = \begin{cases} 1 & \text{if } t \leq \tau \\ 0 & \text{if } t > \tau, \end{cases}$$

where τ corresponds to the cutoff (in days) determining a movie's early life relative to its release date. For instance, a value of $\tau = 7$ would consider a movie's early life as the first seven days after its release date. I then expand equation (3) by interacting all the coefficients with the dummy variables $Early_{it}$ and $Late_{it} = 1 - Early_{it}$ in order to obtain distinct coefficients for the early and late periods of a movie's life.[49]

Table A18 presents the results of estimating various of these specifications across the range of τ values $\tau = \{7, 14, 30, 180\}$. Specifications (1) to (4) use the rating score as the dependent variable and all control for movie and reviewer fixed effects as well as sequential and temporal dynamics. Across all cutoff values, the results show that the difference in the gender score gap across crowds and critics realizes both in the early and late life of movies. As indicated in the bottom row, a t-test on the equality of both the early and late difference-in-differences estimates reveals no statistical difference between the two. Specifications (4)-(8) reveal a similar pattern when focusing on extreme low ratings. Overall, these results are again consistent with herding behavior among reviewers and suggest that reviews posted later in a movie's life contain limited additional information.[50]

[48]Note that given the lack of appropriate instruments for these lagged variables, there is a possibility that including them would bias the difference-in-differences coefficient of interest if these new variables are correlated with the error term. I therefore refrain from including these controls in the main model (3). The

Table A17: Gender Gaps between Critics and Rotten Tomatoes Crowds. †

Dependent Variable:	Rating Score		Extreme Low Rating	
	(1)	(2)	(3)	(4)
	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.
Female Movie × Crowd Reviewer	-1.785***	-1.401***	0.906***	0.637***
	(0.297)	(0.251)	(0.125)	(0.108)
Elapsed Time since First Review	0.711***	0.983***	0.235**	-0.007
	(0.249)	(0.209)	(0.117)	(0.092)
Review Position within Sequence	-0.052***	-0.024***	0.010***	0.006***
	(0.010)	(0.009)	(0.002)	(0.001)
Review Year	-0.575***	-0.349***	0.039	0.001
	(0.084)	(0.068)	(0.037)	(0.030)
Lag Mean Rating Score		0.719***		
		(0.027)		
Lag S.D. of Rating Score		-0.549***		
		(0.038)		
Lag Mean Extreme Low Ratings				1.047***
				(0.043)
Lag S.D. Extreme Low Ratings				-0.017
				(0.012)
Movie Fixed Effects	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓
R^2	0.457	0.461	0.262	0.266
Number of Observations	15825513	15815566	15825513	15815566

† A movie is considered female if the top actress is female. Specifications (1)-(4) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (5)-(8) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. Box office revenue is measured in million USD. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A18: Differences in Gender Score Gaps between Critics and Rotten Tomatoes Crowds. Early vs Late Periods. †

Dependent Variable:	Rating Score				Extreme Low Ratings			
Early/Late Period Cut-off (in days):	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer × Early	-1.244***	-1.340***	-1.403***	-1.503***	0.765***	0.828***	0.860***	0.876***
	(0.388)	(0.389)	(0.384)	(0.339)	(0.168)	(0.167)	(0.163)	(0.141)
Female Movie × Crowd Reviewer × Late	-1.760***	-1.656***	-1.653***	-1.760***	0.803***	0.745***	0.709***	0.638***
	(0.276)	(0.270)	(0.267)	(0.294)	(0.120)	(0.119)	(0.118)	(0.148)
Review Position within Sequence × Early	-0.050***	-0.058***	-0.054***	-0.040***	0.006***	0.006***	0.006***	0.006***
	(0.013)	(0.011)	(0.009)	(0.008)	(0.002)	(0.001)	(0.001)	(0.001)
Review Position within Sequence × Late	-0.031***	-0.032**	-0.036*	-0.043***	0.010***	0.012***	0.012***	-0.008*
	(0.010)	(0.013)	(0.020)	(0.008)	(0.002)	(0.003)	(0.003)	(0.004)
Elapsed Time since First Review × Early	-5.899***	-6.504***	-7.429***	-12.510***	1.048***	1.262***	1.379***	3.091***
	(0.717)	(0.701)	(0.749)	(0.870)	(0.286)	(0.279)	(0.263)	(0.318)
Elapsed Time since First Review × Late	0.402*	0.195	0.173	-0.149	0.056	0.002	-0.096	-0.211*
	(0.240)	(0.242)	(0.230)	(0.171)	(0.108)	(0.107)	(0.112)	(0.125)
Review Year × Early	0.093	0.000	-0.228	-0.581***	-0.088	-0.108	-0.062	-0.015
	(0.182)	(0.163)	(0.146)	(0.100)	(0.079)	(0.071)	(0.054)	(0.043)
Review Year × Late	-0.366***	-0.281***	-0.262***	-0.154***	0.049	0.050	0.059	0.043
	(0.085)	(0.085)	(0.076)	(0.060)	(0.037)	(0.037)	(0.039)	(0.044)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
t-test: Diff. between Early and Late periods	2.56	0.94	0.58	0.64	0.06	0.28	0.96	2.07
p-value	0.110	0.332	0.444	0.423	0.806	0.594	0.328	0.150
R^2	0.473	0.477	0.480	0.479	0.276	0.280	0.285	0.287
Number of Observations	15051250	14932159	14940994	15178547	15051250	14932159	14940994	15178547

† A movie is considered female if the top actress is female. Specifications (1)-(4) use the rating score (measured from 10 to 100) as the dependent variable. Specifications (5)-(8) use a dummy equal to 1 (and multiplied by 100) for ratings equal to 10 as the dependent variable. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Box office revenue is measured in million USD. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A19: Differences in Gender Sentiment Gaps between Critics and Rotten Tomatoes Crowds. Early vs Late Periods. VADER Sentiment Analysis. †

Dependent Variable:	Sentiment Score				Extreme Sentiment Score			
Early/Late Period Cut-off (in days):	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer × Early	-0.024***	-0.026***	-0.025***	-0.026***	0.227***	0.219***	0.213***	0.237***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.077)	(0.072)	(0.068)	(0.060)
Female Movie × Crowd Reviewer × Late	-0.024***	-0.022***	-0.022***	-0.026***	0.221***	0.210***	0.225***	0.138
	(0.004)	(0.004)	(0.004)	(0.005)	(0.066)	(0.069)	(0.074)	(0.098)
Review Position within Sequence × Early	-0.000**	-0.001***	-0.001***	-0.001***	0.002	0.001	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Review Position within Sequence × Late	-0.001***	-0.001***	-0.000***	0.000	0.001	0.001	0.002	-0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.002)	(0.002)
Elapsed Time since First Review × Early	-0.031***	-0.033***	-0.044***	-0.163***	0.433**	0.418***	0.362***	0.301***
	(0.010)	(0.010)	(0.010)	(0.013)	(0.173)	(0.152)	(0.122)	(0.082)
Elapsed Time since First Review × Late	-0.012***	-0.015***	-0.014***	-0.011***	0.161***	0.157***	0.116**	0.180**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.051)	(0.054)	(0.058)	(0.088)
Review Year × Early	-0.001	-0.003	-0.006***	-0.004***	-0.055	-0.031	-0.016	0.007
	(0.003)	(0.002)	(0.002)	(0.002)	(0.050)	(0.040)	(0.029)	(0.017)
Review Year × Late	-0.002	-0.000	-0.000	-0.000	0.004	0.007	0.022	0.019
	(0.001)	(0.001)	(0.001)	(0.001)	(0.018)	(0.019)	(0.020)	(0.031)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
t-test: Diff. between Early and Late periods	0.00	0.61	0.48	0.00	0.01	0.01	0.02	1.10
p-value	0.991	0.435	0.490	0.974	0.930	0.907	0.876	0.295
R^2	0.279	0.282	0.284	0.283	0.152	0.154	0.156	0.156
Number of Observations	13974232	13864631	13875259	14099914	13974232	13864631	13875259	14099914

† A movie is considered female if the top actress is female. All specifications rely on the VADER classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A20: Differences in Gender Sentiment Gaps between Critics and Rotten Tomatoes Crowds. Early vs Late Periods. TextBlob Sentiment Analysis. †

Dependent Variable:	Sentiment Score				Extreme Sentiment Score			
Early/Late Period Cut-off (in days):	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer × Early	-0.016***	-0.016***	-0.016***	-0.016***	0.173***	0.135***	0.141***	0.131***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.051)	(0.048)	(0.046)	(0.040)
Female Movie × Crowd Reviewer × Late	-0.013***	-0.012***	-0.012***	-0.014***	0.094**	0.100**	0.071	0.034
	(0.002)	(0.002)	(0.002)	(0.003)	(0.046)	(0.048)	(0.051)	(0.070)
Review Position within Sequence × Early	-0.000***	-0.000***	-0.000***	-0.000***	0.000	0.002*	0.002***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Review Position within Sequence × Late	-0.000***	-0.000***	-0.000***	-0.001***	0.004***	0.005***	0.006***	0.004**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)
Elapsed Time since First Review × Early	-0.024***	-0.024***	-0.028***	-0.062***	0.204**	0.238***	0.323***	0.965***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.097)	(0.087)	(0.081)	(0.087)
Elapsed Time since First Review × Late	-0.004**	-0.005**	-0.004**	-0.004*	0.161***	0.181***	0.169***	0.015
	(0.002)	(0.002)	(0.002)	(0.002)	(0.044)	(0.045)	(0.049)	(0.066)
Review Year × Early	0.001	-0.000	-0.001	-0.003***	-0.021	-0.015	-0.018	-0.022
	(0.001)	(0.001)	(0.001)	(0.001)	(0.031)	(0.026)	(0.022)	(0.016)
Review Year × Late	-0.002***	-0.002**	-0.002***	-0.002**	-0.023	-0.036**	-0.037**	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.015)	(0.016)	(0.017)	(0.024)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
t-test: Diff. between Early and Late periods	1.08	2.53	2.16	0.38	1.48	0.30	1.17	1.59
p-value	0.300	0.112	0.141	0.536	0.224	0.586	0.279	0.208
R^2	0.297	0.299	0.303	0.303	0.181	0.184	0.188	0.192
Number of Observations	13974232	13864631	13875259	14099914	13974232	13864631	13875259	14099914

† A movie is considered female if the top actress is female. All specifications rely on the TextBlob classifier to construct reviews' sentiment score. The score ranges from -1 to 1, with a higher score reflecting a more positive sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 1^{st} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A21: Differences in Gender Sentiment Gaps between Critics and Rotten Tomatoes Crowds. Early vs Late Periods. LIWC Sentiment Analysis. †

Dependent Variable:	Sentiment Score				Extreme Sentiment Score			
Early/Late Period Cut-off (in days):	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$	$\tau = 7$	$\tau = 14$	$\tau = 30$	$\tau = 180$
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie × Crowd Reviewer × Early	0.098**	0.107***	0.093**	0.119***	0.156***	0.157***	0.144***	0.149***
	(0.042)	(0.041)	(0.039)	(0.034)	(0.051)	(0.049)	(0.046)	(0.040)
Female Movie × Crowd Reviewer × Late	0.134***	0.125***	0.141***	0.128***	0.134***	0.126**	0.134**	0.103
	(0.033)	(0.034)	(0.036)	(0.047)	(0.048)	(0.051)	(0.055)	(0.075)
Review Position within Sequence × Early	0.002***	0.003***	0.003***	0.002***	0.001	0.003***	0.003***	0.002***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)
Review Position within Sequence × Late	0.002**	0.002**	0.003	0.004**	0.002**	0.003	0.002	0.007***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.004)	(0.002)
Elapsed Time since First Review × Early	0.375***	0.387***	0.463***	1.038***	0.331***	0.314***	0.476***	1.213***
	(0.069)	(0.063)	(0.061)	(0.069)	(0.116)	(0.105)	(0.095)	(0.099)
Elapsed Time since First Review × Late	0.111***	0.136***	0.130***	0.061	0.241***	0.270***	0.278***	0.144*
	(0.028)	(0.030)	(0.033)	(0.039)	(0.050)	(0.054)	(0.062)	(0.077)
Review Year × Early	-0.025	-0.013	-0.004	0.001	-0.022	-0.010	-0.033	-0.020
	(0.021)	(0.018)	(0.016)	(0.012)	(0.035)	(0.030)	(0.024)	(0.018)
Review Year × Late	-0.004	-0.016	-0.017	-0.006	-0.025	-0.039**	-0.045**	-0.027
	(0.010)	(0.010)	(0.011)	(0.014)	(0.018)	(0.019)	(0.021)	(0.027)
Movie Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Reviewer Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
t-test: Diff. between Early and Late periods	0.65	0.14	1.10	0.03	0.10	0.21	0.02	0.30
p-value	0.421	0.705	0.295	0.852	0.750	0.643	0.885	0.584
R^2	0.229	0.232	0.235	0.237	0.182	0.184	0.188	0.191
Number of Observations	13972288	13862698	13873308	14097982	13972288	13862698	13873308	14097982

† A movie is considered female if the top actress is female. All specifications rely on the LIWC's negative tone dictionary to construct reviews' sentiment score, ranging from 0 to 100. A higher score reflects a more negative review sentiment. Columns (1)-(4) use the review sentiment score as the dependent variable. Columns (5)-(8) use a dummy equal to 1 (and multiplied by 100) if the review's sentiment score is extreme as the dependent variable. Extreme scores are defined as scores that fall in the 99^{th} percentile of the review sentiment score distribution. The variables corresponding to the Elapsed Time since First Review and the Review Position within Sequence are divided by 1,000 to facilitate the readability of the corresponding coefficients. Crowd ratings come from Rotten Tomatoes. Standard errors are clustered at the movie level and reported in parenthesis.

* Significant at the 10% level.
** Significant at the 5% level.
*** Significant at the 1% level.

Estimation Relying on Cross-Movie Variation Given the large number of crowd reviews posted for each movie and the results presented above, one alternative empirical approach is to construct, for each movie, a measure of the average rating score and conduct a cross-sectional analysis at the movie level rather than at the review level. Beyond providing a robustness check of the results presented above, such approach has several advantages. First, it does not need to address dynamic concerns. Second, it also reduces concerns about the clustering of standard errors given that all the data is collapsed at the movie level. Finally, such approach can also be implemented using the IMDb crowd data and therefore offers another way of comparing the results across the two platforms. I therefore estimate specifications of the following form:

$$\Delta_m = \quad \alpha + \delta FemaleMovie_m + \beta X_m + \varepsilon_m, \tag{1.a}$$

where $\Delta_m \equiv \overline{score}_m^{crowd} - \overline{score}_m^{critics}$ is the difference in the average score received by crowd reviewers and critics for movie m. The coefficient δ therefore directly corresponds to the difference-in-difference estimate of interest.

Column (1) of Table A22 report the results of estimating equation (1.a) using the Rotten Tomatoes crowd data. Consistent with the results presented above, the difference-in-differences estimates is equal to -1.9, almost identical to our preferred estimate in specification (4) in Table 4. Using the IMDb crowds data – in specification (3) – leads to a similar estimate. Specification (2) and (4) perform similar exercises using the difference in the share of extreme low ratings received by crowd reviewers and critics by each movie. Perhaps unsurprisingly, the estimates are again consistent with the results presented above and show that the gender gap in extreme low ratings is significantly larger among crowds than critics, and even more so on Rotten Tomatoes than IMDb.[51] Table A23 reports the results of estimating similar specifications using sentiment scores as dependent variables. All show similar results.

results presented in Table A17 nevertheless provide a good test for the robustness of the main results.

[49]In particular, I also interact the movie and reviewer fixed effects with the dummy variable for the early and late periods. Note that such empirical specification is equivalent to splitting the sample between the early and late periods. However, relying on the full sample has the advantage of allowing to test whether the coefficients of interest are statistically different across periods.

[50]Tables A19, A20, and A21 present the results of performing similar estimations relying on the VADER, TextBlob, and LIWC sentiment scores as dependent variables, respectively. All show similar results as the ones presented in Table A18.

[51]Tables A24 and A25 present the corresponding estimation results when relying on less restrictive definitions of female movies using the Rotten Tomatoes crowd data and IMDb crowd data, respectively. The results are once again robust to such different definitions.

Table A22: Gender Gaps between Critics and Crowd using Cross-Section of Movies. †

	Rotten Tomatoes Crowds		IMDb Crowds	
Dependent Variable Based On:	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.
Female Movie	-1.890***	0.853***	-1.550***	0.419**
	(0.295)	(0.258)	(0.225)	(0.195)
Worldwide Box Office Revenue	0.003***	-0.001**	-0.000	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)
Movie Released in Theaters	0.800*	-0.979**	0.883***	-1.443***
	(0.415)	(0.397)	(0.318)	(0.297)
Number of Release Countries	-0.010	-0.022***	0.028***	-0.058***
	(0.010)	(0.007)	(0.008)	(0.005)
Number of Awards Won	-0.133***	-0.002	-0.138***	-0.001
	(0.009)	(0.005)	(0.009)	(0.003)
Genre Fixed Effects	✓	✓	✓	✓
Production Studio Fixed Effects	✓	✓	✓	✓
MPAA Rating Fixed Effects	✓	✓	✓	✓
Year of Release Fixed Effects	✓	✓	✓	✓
Origin Fixed Effects	✓	✓	✓	✓
Language Fixed Effects	✓	✓	✓	✓
R^2	0.110	0.086	0.158	0.084
Number of Observations	12657	12657	12657	12657

† A movie is considered female if the top actress is female. Each observation in the sample corresponds to a movie. In specifications (1) and (3), the dependent variable is the difference in the average rating score between crowd and critics. In specifications (2) and (4), the dependent variable is the difference in the share of extreme low ratings between crowd and critics. Box office revenue is measured in million USD. Robust standard errors are reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A23: Gender Sentiment Score Gaps between Critics and Crowd using Cross-Section of Movies. †

	Sentiment Score			Extreme Sentiment Score		
Sentiment Classifier:	VADER	TextBlob	LIWC	VADER	TextBlob	LIWC
	(1)	(2)	(3)	(4)	(5)	(6)
	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.	Coef./s.e.
Female Movie	-0.035***	0.182	-0.016***	0.172**	0.224***	0.271**
	(0.006)	(0.116)	(0.003)	(0.081)	(0.063)	(0.132)
Worldwide Box Office Revenue	-0.000	-0.000	0.000***	-0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Movie Released in Theaters	0.014	-0.392**	0.012**	-0.010	0.054	0.109
	(0.009)	(0.181)	(0.005)	(0.137)	(0.101)	(0.218)
Number of Release Countries	0.001***	-0.014***	0.000***	0.002	0.002	0.007*
	(0.000)	(0.002)	(0.000)	(0.002)	(0.002)	(0.004)
Number of Awards Won	-0.000	-0.003	-0.000***	-0.003**	0.001	0.004**
	(0.000)	(0.002)	(0.000)	(0.001)	(0.001)	(0.002)
Genre Fixed Effects	✓	✓	✓	✓	✓	✓
Production Studio Fixed Effects	✓	✓	✓	✓	✓	✓
MPAA Rating Fixed Effects	✓	✓	✓	✓	✓	✓
Year of Release Fixed Effects	✓	✓	✓	✓	✓	✓
Origin Fixed Effects	✓	✓	✓	✓	✓	✓
Language Fixed Effects	✓	✓	✓	✓	✓	✓
R^2	0.041	0.057	0.067	0.023	0.021	0.019
Number of Observations	12566	12657	12566	12657	12566	12657

† A movie is considered female if the top actress is female. Each observation in the sample corresponds to a movie. In specifications (1) and (3), the dependent variable is the difference in the average sentiment score between crowd and critics. In specifications (2) and (4), the dependent variable is the difference in the share of extreme sentiment score between crowd and critics. Box office revenue is measured in million USD. Crowd review scores come from Rotten Tomatoes. Robust standard errors are reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A24: Gender Gaps between Critics and Rotten Tomatoes Crowd using Cross-Section of Movies. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie	-1.890***	0.853***	-1.386***	0.820***	-1.202***	0.679**	0.336	0.188
	(0.295)	(0.258)	(0.292)	(0.259)	(0.351)	(0.320)	(0.392)	(0.230)
Worldwide Box Office Revenue	0.003***	-0.001**	0.003***	-0.001**	0.003***	-0.001**	0.002**	-0.001***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Movie Released in Theaters	0.800*	-0.979**	0.831**	-0.992**	0.834**	-0.994**	-0.207	-0.042
	(0.415)	(0.397)	(0.416)	(0.397)	(0.416)	(0.397)	(0.969)	(0.701)
Number of Release Countries	-0.010	-0.022***	-0.011	-0.022***	-0.011	-0.022***	0.023*	-0.010
	(0.010)	(0.007)	(0.010)	(0.007)	(0.010)	(0.007)	(0.013)	(0.008)
Number of Awards Won	-0.133***	-0.002	-0.133***	-0.002	-0.133***	-0.002	-0.120***	0.005
	(0.009)	(0.005)	(0.009)	(0.005)	(0.009)	(0.005)	(0.008)	(0.004)
Genre Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Production Studio Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
MPAA Rating Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Year of Release Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Origin Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Language Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.110	0.086	0.108	0.086	0.108	0.085	0.224	0.193
Number of Observations	12657	12657	12657	12657	12657	12657	3571	3571

† Each observation in the sample corresponds to a movie. For each Female movie definition, the first specification uses the difference in the average rating score between crowd and critics as the dependent variable, while the second uses the difference in the share of extreme low ratings between crowd and critics as the dependent variable. Crowd ratings come from Rotten Tomatoes. Box office revenue is measured in million USD. Robust standard errors are reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table A25: Gender Gaps between Critics and IMDb Crowd using Cross-Section of Movies. †

Female Movie Definition Based On:	Female Lead		Female in Top 2		Female in Top 3		Bechdel Test	
Dependent Variable:	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating	Rating Score	Extreme Low Rating
	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.	(7) Coef./s.e.	(8) Coef./s.e.
Female Movie	-1.550***	0.419**	-1.037***	0.405*	-0.762***	0.233	-0.194	0.465***
	(0.225)	(0.195)	(0.228)	(0.208)	(0.281)	(0.267)	(0.322)	(0.175)
Worldwide Box Office Revenue	-0.000	0.002***	-0.000	0.002***	-0.000	0.002***	-0.000	0.001***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Movie Released in Theaters	0.883***	-1.443***	0.908***	-1.450***	0.910***	-1.450***	0.550	-0.580
	(0.318)	(0.297)	(0.319)	(0.297)	(0.319)	(0.297)	(0.784)	(0.573)
Number of Release Countries	0.028***	-0.058***	0.027***	-0.058***	0.026***	-0.058***	0.042***	-0.034***
	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.010)	(0.005)
Number of Awards Won	-0.138***	-0.001	-0.139***	-0.001	-0.139***	-0.001	-0.115***	0.001
	(0.009)	(0.003)	(0.009)	(0.003)	(0.009)	(0.003)	(0.008)	(0.002)
Genre Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Production Studio Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
MPAA Rating Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Year of Release Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Origin Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
Language Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.158	0.084	0.157	0.084	0.156	0.084	0.270	0.125
Number of Observations	12657	12657	12657	12657	12657	12657	3571	3571

† Each observation in the sample corresponds to a movie. For each Female movie definition, the first specification uses the difference in the average rating score between crowd and critics as the dependent variable, while the second uses the difference in the share of extreme low ratings between crowd and critics as the dependent variable. Crowd ratings come from Rotten Tomatoes. Box office revenue is measured in million USD. Robust standard errors are reported in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.