How Did This Get Rated?

Media Persuasion and the Formation of Online Ratings*

Benjamin T. Leyden†

November 3, 2026

Abstract

I study how media coverage—specifically, popular movie podcasts—affects movie ratings using data from two popular platforms, IMDb and Letterboxd. Using a stacked difference-in-differences design, I estimate that podcast coverage increases the number of ratings a movie receives and lowers its average rating. Moreover, I find that the negative effect on ratings is stronger the more negatively a show's hosts assess the film. Using individual-level Letterboxd ratings, I show that this effect is driven by the persuasive effect of the hosts' discourse and not only by a change in who rates. This provides evidence that the content of media coverage plays a role in shaping consumers' perception of products, and of the reputation of those products online.

JEL Classification: L15, L82, L86.

Keywords: rating systems, online platforms, information intermediaries, media coverage.

*I thank Yiting Deng and Felix Schleef for helpful comments and suggestions.

†SC Johnson College of Business, Cornell University; CESifo; leyden@cornell.edu.

1 Introduction

Online rating systems have become an essential part of digital markets as a result of the often substantial information asymmetries between buyers and sellers (Dellarocas, 2003; Tadelis, 2016). These systems aggregate dispersed consumer assessments into summary statistics that guide purchase decisions (Chevalier and Mayzlin, 2006; Anderson and Magruder, 2012; Luca, 2016; Reimers and Waldfogel, 2021) and shape firm incentives (Leyden, 2025). Rating systems have also been shown to be susceptible to bias (Dellarocas and Wood, 2008; Nosko and Tadelis, 2026; Filippas, Horton, and Golden, 2022; Luca and Zervas, 2016; He, Hollenbeck, and Proserpio, 2022; Mayzlin, Dover, and Chevalier, 2014; Zuloaga Cosme, Parker, and Beck, 2025; Kovács and Sharkey, 2014). In this paper, I consider another source of influence on product ratings: media coverage. Specifically, I ask whether external media coverage affects who shows up to rate and how they rate.

I study how two popular movie podcasts (*Blank Check* and *How Did This Get Made?* (*HDTGM*)) drive listeners to rate established *catalog films* (which I define as those released at least 2 years prior with at least 1,000 existing votes). Podcasts are a natural setting for studying whether and how media coverage changes the ratings people give. Their long-form, conversational format gives hosts room to develop and debate their ideas, and listeners who follow the same hosts week after week often develop a parasocial relationship à la Horton and Wohl (1956) that can amplify the persuasive impact of a host's opinion relative to written reviews or algorithmic recommendations. The two shows share a similar structure, with each devoting an episode to a single film, although their purpose and tone differ. *Blank Check*'s hosts and guests conduct deep-dives on the work of successful directors (those who "are given a series of blank checks to make whatever crazy passion projects they want"), while *HDTGM*'s celebrate films they consider bad.

I track rating activity on IMDb and Letterboxd, two dominant platforms for movie ratings and reviews. Using a stacked difference-in-differences (DiD) design, I measure the effect of podcast coverage on how many people rate (the "extensive margin") and how they rate (the "intensive margin"). I then score episode transcripts on recommendation valence and quality assessment to estimate the effect of what the hosts say about a movie on its ratings.

I find that coverage raises the volume of new ratings on both platforms by 16.4% on IMDb and 127.9% on Letterboxd. Coverage induces a sharp increase in rating activity that begins in the week

before air and peaks around the air date, after which it fades over the course of roughly two weeks. Movies' average ratings fall as a result, and the fall deepens with how negatively the hosts assess the film: across films, more-panned films lose more. That finding, however, is consistent both with the podcast changing its listeners' opinions and with its drawing a more critical crowd. To address this, I leverage user-level ratings data from Letterboxd to separate them. Holding the reviewer fixed, I find that the same person rates a more-panned film lower and a more-praised one higher: moving the hosts' assessment across its observed range moves the same reviewer's rating by 0.23 stars—over a quarter of the typical spread in a single reviewer's ratings. Thus, media coverage changes not only who rates but how a given person rates.

The composition of online ratings is driven by who chooses to rate a product and why. Self-selection produces systematically non-random rating populations: early adopters differ from late adopters (Li and Hitt, 2008), extreme-opinion consumers are more likely to review (Hu, Pavlou, and Zhang, 2017; Schoenmueller, Netzer, and Stahl, 2020), and social cues shift rating behavior (Muchnik, Aral, and Taylor, 2013; Moe and Trusov, 2011; Moe and Schweidel, 2012; Chen et al., 2010). To the extent that external actors shift the composition of the rater pool or the preferences of those raters, they change what the platform metric reflects, even though the underlying product itself has not been altered.

Recent work has shown that external events can shift demand and online reviews. On the demand side, Garthwaite (2014) shows that Oprah's Book Club endorsements created large demand spillovers, Berger, Sorensen, and Rasmussen (2010) find that negative publicity improves sales for lesser-known products, suggesting awareness can dominate valence, and Brandes, Nüesch, and Franck (2016) show that an artist's death can serve as informational advertising, increasing music sales. Most directly related to this paper, Kovács and Sharkey (2014) and Rossi and Schleef (2026) show that prestigious awards lead to lower average ratings. In the context of Goodreads book reviews, Kovács and Sharkey (2014) find that winning a prominent award both draws a more diverse set of reviewers and prompts a negative reaction to the book's increased popularity. In the context of movie ratings, Rossi and Schleef (2026) show that Oscar nominations lead to a decline in MovieLens ratings. They attribute this decline to consumer disappointment, as a result of the award nomination leading to elevated expectations, rather than to a change in who rates the film, and this effect seems largest among the platform's least experienced users. In a similar vein, Li

et al. (2026) find that when a restaurant loses a Michelin star, its TripAdvisor ratings improve, which they attribute to lowered expectations.

In addition to drawing attention to a film, a podcast episode includes an extended discussion of its content and the hosts' evaluation of its quality. This relates to a body of work about how information provision for experience goods shapes demand and product evaluation. Evidence is mixed on whether critics influence demand, as opposed to simply predicting it. Eliashberg and Shugan (1997) find that critics tend to act more as predictors than influencers of ticket sales, while Reinstein and Snyder (2005) provide evidence that critics' reviews affect demand for dramas and long-tail, niche films. Reimers and Waldfogel (2021) find that a *New York Times* review raises a book's sales, and more so when the *Times* also recommends the book. Beyond the scope of professional critics, there is evidence that social interactions affect movie demand (Moretti, 2011; Gilchrist and Sands, 2016), as does word-of-mouth (Liu, 2006). In closely related work, Aridor et al. (2026) use a field experiment on MovieLens to show that recommendations raise consumption through an informational channel in addition to increased exposure. In earlier work on the same platform Cosley et al. (2003) show that recommendations can move the rating itself: shifting the predicted MovieLens rating displayed to users leads users to adjust their rating toward the displayed prediction. I build on this work by directly measuring the content of the coverage from episode transcripts and, in the final step of my analysis, by following the same Letterboxd reviewer across films, which allows me to estimate how the content of the podcasts affects how a person rates a film.

I contribute to our understanding of the formation of online ratings in three primary ways. First, I use individual ratings to separate persuasion from selection. The film-level average rating cannot distinguish a fixed group of raters who change their minds from a different group of raters who show up, since both move the average in the same direction. By holding the reviewer fixed, I am able to estimate how the content of the podcast affects how a given person rates a film, separately from its effect on who rates. Second, I measure what the podcasts say about each film directly from episode transcripts and link these measures to each margin of rating activity. I find that coverage raises the volume of new ratings whether the hosts praise or pan the film, while the rating a film receives follows the hosts' assessment of its quality. Finally, I show that the inputs to a platform's rating are endogenous to the surrounding media environment. Media coverage shifts

both the volume and the composition of the ratings a major platform aggregates, which matters alongside the aggregation rule itself (Dellarocas, 2003; Leyden, 2025).

The remainder of the paper proceeds as follows. In Section 2, I present a conceptual framework of the three channels—attention, persuasion, and selection—through which media coverage can change who rates a film and how they rate it. In Section 3, I describe the setting and data. In Section 4, I present the empirical analysis, and I summarize a set of robustness checks in Section 5. Finally, in Section 6, I discuss the implications of these results and conclude.

2 Conceptual Framework

This section develops a framework for how external attention intermediaries, such as movie podcasts, can shift the composition and content of online ratings. Coverage can affect both how many people rate (the extensive margin) and what ratings they give (the intensive margin), and I focus on three channels through which it does so: attention, persuasion, and selection. The framework generates three testable predictions. The first two concern the film-level average rating, on which—as I show below—persuasion and selection cannot be separated; the third concerns the individual reviewer, on which they can.

2.1 Setup

Consider a consumer i who is deciding whether to watch and rate movie j, which is covered by podcast episode e. Consumer i watches movie j when the expected utility of watching, net of its opportunity cost, is positive:

$$W_i(j, e) \;=\; \mathbb{E}_i\big[u(q_{ij})\big] - c_i \;>\; 0, \tag{1}$$

where q_{ij} is i's realized quality assessment and c_i is the opportunity cost of watching. The episode e—a podcast's discussion of film j—is the treatment. A podcast episode can raise W_i by increasing expected utility through improved information about j (shifting the expectation of q_{ij} and thus the watch decision), or by lowering the effective opportunity cost of watching (by increasing the salience of the movie and/or the social return to watching the movie). This latter case includes the parasocial

benefits of participating in the podcast's community (e.g., engaging in online discussions), which makes watching valuable independent of the actual quality of the film (Seiler, Yao, and Wang, 2017).

The assessment q_{ij} is drawn from a distribution $F(\cdot \mid j)$, and movie j has an established average rating $\bar{r}_j$ based on N_j prior votes. Because j is a catalog film—one released at least 2 years prior with at least 1,000 existing votes—quality beliefs are settled and I assume baseline rating activity is stable. With $F(\cdot \mid j)$ stationary, I can attribute any shift in rating behavior following podcast coverage to the podcast rather than to evolving beliefs about the film.[1]

Conditional on watching, i rates movie j if the benefit of doing so exceeds a cost k_i. The benefit decomposes as

$$B_i^{\text{rate}}(j, e) \;=\; b_i^{\text{expr}}(q_{ij}) \;+\; b_i^{\text{social}}(j, e) \;+\; b_i^{\text{info}}(j, N_j), \tag{2}$$

where b_i^{expr} is expression utility (the intrinsic value of registering an opinion), b_i^{social} is community utility (participating in a shared experience amplified by the podcast community), and b_i^{info} is informational utility (contributing to the public good $\bar{r}_j$), which is likely small for catalog films with large N_j. The podcast raises b_i^{social} and b_i^{expr}: listeners anticipate discussing the film in the podcast's community, and the episode primes the act of evaluation.

The rating i submits for j is

$$r_{ij} \;=\; q_{ij}. \tag{3}$$

This is the avenue through which persuasion can occur. The podcast hosts/guests' discussion of the film can change the assessment the viewer forms after watching, not whether they watch. The consumer's decision to watch film j (Equation (1)) depends on the expected quality of the movie, while the rating (Equation (3)) reflects its realization. The average rating $\bar{r}_j$ will shift to the extent that the podcast shapes the distribution of q_{ij} across mobilized raters, and the three channels differ in whether, and how, they do so.

[1] Catalog films can undergo critical reappraisal over time—generational shifts in taste, rediscovery, or retrospective reassessment. This will only present a threat to the analysis here if these reappraisals are systematically correlated with the timing of podcast coverage. This is unlikely given the way in which the two shows pick movies and also in light of the matched-control design.

2.2 Attention, Persuasion, and Selection

I now consider each of the three channels in turn. Attention makes film j salient, which raises the chance that a consumer watches it (Equation (1)) and, having watched it, rates it (Equation (2)). Consumers who rate for this reason still draw q_{ij} from $F(\cdot|j)$, so attention on its own—henceforth, "pure attention"—predicts that more people will rate, but the average rating will not change. Persuasion changes the rating itself (Equation (3)) because the hosts' discussion moves a viewer's assessment q_{ij} closer to the hosts' own opinion of the film, which I summarize by the episode's transcript-scored quality assessment $s_e \in [-1, 1]$. If the hosts are more positive about the film than the existing raters, q_{ij} will rise relative to F, whereas if they are more negative, q_{ij} will fall. Thus, persuasion predicts that the average rating moves in the direction of what the hosts say.

Selection, on the other hand, changes *who* rates the movie. To the extent that the podcast spurs a non-random subset of F to watch and rate (Equation (1)), and that subset (the show's listeners in general, or of a particular episode) holds a different opinion of the film than the existing raters do, the average rating moves as well (Hu, Pavlou, and Zhang, 2017). Selection can therefore move the average rating in the same direction as persuasion, but it does so by changing the composition of the raters rather than the opinion of any one of them.

Given these channels, I propose three testable predictions, two at the film level and one at the individual reviewer level.

First, all three channels predict an increase in the number of ratings. Podcast coverage raises W_i and B_i^{rate}, increasing both the probability of watching and the probability of rating conditional on watching. Coverage raises volume even for films the hosts dislike because attention and social utility operate through the salience of the film rather than through its quality signal. This leads to the framework's first prediction:

Prediction 1 (Extensive-margin response)**.** *Podcast coverage will increase the number of ratings a film receives, regardless of whether the hosts think the film is good or bad.*

Under pure attention, podcast-mobilized raters draw q_{ij} from $F(\cdot|j)$—the same distribution as existing raters—so $\bar{r}_j$ does not shift. Persuasion will move $\bar{r}_j$ in the direction of the episode's quality assessment, because each mobilized rater's q_{ij} moves toward the hosts' opinion. Selection can produce the same pattern if the raters an episode mobilizes share the hosts' tastes and can

produce the reverse pattern if their tastes are unrelated or counter to the hosts' assessment.

Prediction 2 (Intensive-margin response)**.** *Under pure attention, a film's average rating $\bar{r}_j$ will not change. Under persuasion, $\bar{r}_j$ will move in the direction of the hosts' assessment of it, falling when the hosts are negative about the film and rising when they are positive. Selection can produce the same pattern, attenuate it, or reverse it, depending on how the tastes of the raters an episode mobilizes relate to the hosts' assessment.*

Importantly, the effects of persuasion versus selection on the film-level average can be observationally equivalent. The individual-level Letterboxd ratings allow me to analyze the effect of podcast coverage on ratings while holding the rater's identity fixed, which helps separate the two effects. Consider a reviewer i who rates several covered films, each discussed in an episode with its own quality assessment s_e. Persuasion changes the same reviewer's rating: i rates a more-panned covered film lower and a more-praised one higher, so i's rating tracks s_e even after conditioning on i. Let θ denote the within-reviewer slope of the rating in s_e.

Prediction 3 (Within-reviewer response)**.** *Under persuasion, the same reviewer will rate a covered film higher when the hosts are more positive about it and lower when they are more negative, so $\theta > 0$. If selection sorts reviewers only by how harshly or generously they rate, and not by how closely their tastes track the hosts', the reviewer's rating will not depend on the hosts' assessment, so $\theta = 0$.*

Selection can still produce $\theta > 0$ if the reviewers a show mobilizes share its hosts' tastes more than the film's existing reviewers do, since holding the reviewer fixed removes the level of their ratings but not how closely their tastes track the hosts'. I return to this in Section 4.5.

3 Setting and Data

I next introduce the two podcasts and then describe the two movie rating platforms, the construction of the analysis sample, and how I score transcripts to measure what the podcast hosts and guests say about each film. My analysis covers podcast episodes airing between December 2015 and December 2025.

3.1 Podcasts as Attention Intermediaries

Movie podcasts are long-form audio programs in which hosts discuss individual films at length. Listeners hear an episode on demand, mostly in the days and weeks after its release, rather than at a scheduled time. The shows are conversational in nature, with hosts making small talk, sharing opinions, and recommending (or discouraging) viewing. Listeners often develop parasocial relationships with hosts—the quasi-intimacy that Horton and Wohl (1956) describe as characteristic of mass-mediated communication. Hosts often develop recurring bits and tell stories about their own lives, sharing details about their families, their careers, and the circumstances in which they first saw a film, so that a regular listener comes to feel that they know the hosts personally. This can amplify the persuasive weight of host opinions relative to written reviews or algorithmic recommendations.

I study episodes from two popular movie podcasts: 356 *Blank Check* episodes airing between December 2015 and November 2025, and 212 *HDTGM* episodes airing between December 2022 and December 2025.[2] *Blank Check* follows directors' complete filmographies chronologically, hits and misses alike. *How Did This Get Made?* (*HDTGM*) is a celebration of films that are, in the hosts' phrasing, "so bad they're good" (Scheer, Raphael, and Mantzoukas, 2012). In a representative episode on *The Garbage Pail Kids Movie* (1987), a guest captures the ethos: "This is the worst movie I have ever seen. It is irredeemable. It is insane... I think people should see it. I think people should check it out" (Scheer, Raphael, and Mantzoukas, 2017).

The two shows differ in how they select films, but in both cases the film is typically announced publicly before the episode is released. *Blank Check* is organized around "miniseries" of weekly episodes, which cover the filmography of a prominent director in release order. The hosts typically choose which directors to cover, although listeners are given the opportunity to select the director for one miniseries each year through a "March Madness" style bracket competition. The director for each miniseries is announced before it begins, and listeners know in advance which film each future episode will cover. *HDTGM*'s film selection is more idiosyncratic, with the hosts choosing which films to cover episode by episode. *HDTGM* releases full episodes biweekly. In the intervening weeks, the show releases a shorter "Last Looks" episode, which engages with listener feedback and

[2] *HDTGM* has aired since December 2010, but its public RSS feed carries only its more recent episodes, so the sample includes its episodes from December 2022 onward. *Blank Check*'s archive is public back to the show's 2015 launch.

announces the next film the show will cover. In both cases, listeners are typically aware of which film will be covered next and can watch it in anticipation of listening to the episode if they wish. Indeed, an informal 2023 poll of the *Blank Check* subreddit suggests that many of them do: 33% of respondents report always watching the film before listening to the episode and only 25% usually listen before having seen it, with the rest split or without a fixed habit (Troile, 2023).

3.2 Movie Rating Platforms

I collect data from two popular movie rating platforms. The first is IMDb, the dominant online database for film information, with over 200 million unique monthly visitors (IMDb, 2026b). Each movie has a weighted average rating on a 1–10 scale, rounded to the nearest tenth of a point. The weighting scheme is proprietary and designed to limit ballot-stuffing.[3]

I observe each film's cumulative vote count and displayed average rating over time. Historical ratings and vote counts are obtained from the Historical IMDb Ratings service, which provides a longitudinal record of IMDb title pages (Historical IMDb Ratings, 2026). Through January 20, 2023, the service reconstructs this record from archived copies of the title pages in the Internet Archive's Wayback Machine, and the resulting snapshots are sparse. Beginning on January 21, 2023, the service records each title page daily.[4] I collect film metadata such as genre and release year from the IMDb non-commercial datasets (IMDb, 2026a). IMDb does not publicly report daily new votes, so I construct the daily flow of new ratings as the first difference of the cumulative vote count.[5]

The second platform is Letterboxd, a social platform for logging and reviewing films, where users rate films on a half-star scale from 0.5 to 5. Unlike on IMDb, where I observe only an average rating for each film, Letterboxd displays the individual ratings behind that average. This makes two things possible that the IMDb data does not. First, it allows me to calculate the daily mean of the film's *new* ratings, allowing me to observe new ratings even for films with long rating histories,

[3] IMDb states that it uses a "weighted average," not the "arithmetic mean," and that it may apply a different weighting when it detects "unusual rating activity" (IMDb, 2026c). The exact weights are undisclosed. Given catalog films typically have large baseline vote counts, new votes will move the displayed average by smaller amounts than the 0.1 display precision, which limits the magnitude of intensive-margin effects the data can address.

[4] Of the 568 treatment events, 249 air before January 21, 2023 and 319 after. In Section 5.2, I show that restricting to the full-coverage (i.e., post-January 21, 2023) IMDb data does not affect the results.

[5] In cases where consecutive snapshots are more than one day apart, which can occur in the sparse, pre-2023 data, I divide the difference by the number of days between them, giving me a daily average over that gap.

when the aggregate score may not move quickly. Second, because each rating carries a user identity, I can follow the same reviewer across the films they rate, which allows me to test Prediction 3.

I collect the reviews of each treated and control film whose diary date—the day the reviewer logs having watched the film, which is not necessarily the day the review is posted—falls within the 30-day event window around the covering episode. This is the most appropriate date for the volume outcome, since it identifies when the *viewing* happened rather than when the review was recorded. After excluding films with fewer than 5 pre-coverage reviews and the truncated-window films described below, the estimation sample contains 130,003 reviews written by 72,750 reviewers, for 261 treated and 739 control films.

There are two limitations worth noting. First, the ratings I observe are those associated with written reviews, so they capture a subset of Letterboxd users. Letterboxd has a proud reviewing tradition: reviews are often short and pithy, and the population writing them is one of movie buffs with strong, discriminating opinions. They are thus arguably harder to sway than the average viewer, and so the within-reviewer responses I estimate will likely understate what a broader audience would show. Second, a film's review listing online is traversable only to a fixed depth, so for the films that accumulate reviews fastest, the event window cannot be fully recovered. As a result, the films in the Letterboxd sample skew toward lower-velocity titles.[6]

3.3 Sample Construction

I define the event window as 30 days before and after each episode's air date. I focus my analysis on films with settled reputations and stable baseline rating trajectories, which I term *catalog films*. Studios and streaming platforms use "catalog" or "library" loosely to describe older titles, though the terms do not have a standardized definition. I define catalog films based on two criteria. First, the film must have been released at least 2 years before the podcast episode airs. This restriction excludes new releases, whose rating activity could be confounded by contemporaneous marketing campaigns and awards-season attention. Second, a film must have at least 1,000 IMDb votes as of

[6]The reviews can be read from either end of the film's review history—newest-first or oldest-first—but only for approximately 3,000 reviews in either direction. A movie's event window is therefore recoverable when it falls within that range. For the most popular films, which accumulate reviews quickly, a window deep in the middle of their histories is out of reach. A covered film with a larger response is thus more likely to be dropped, and so, if anything, this biases my estimate of the volume effect toward zero. I drop films with missing or truncated windows from the Letterboxd analysis.

the last snapshot before the corresponding podcast episode airs, which ensures that in-sample films have a well-established reputation prior to podcast coverage.

A film can potentially be included in the sample more than once for two reasons.[7] First, the same podcast may re-air or revisit a film; I treat that as a single treatment and keep only the show's first coverage. Second, both *Blank Check* and *HDTGM* may cover the same film. I keep each as a separate treatment event since the two different shows reach largely separate audiences, so a second show's coverage is a fresh treatment rather than a repeat of the first. Cross-podcast coverage of this kind spans 8 films, contributing 16 treatment events.[8]

For each treated film, I identify 5 matched control films using three criteria. Control films must have been released within 3 years of the focal film, their baseline vote count must be within 50% of the treated film's baseline count, and they must share at least one IMDb genre with the treated film.[9] The genre requirement makes it more likely that a control film is exposed to the same genre-level shifts in attention as the treated film. I exclude from the control group every film matched to a *Blank Check* or *HDTGM* episode in my data, at any date and whether or not that episode is a treatment event. Given the set of films that meet these criteria, I select the 5 nearest neighbors based on Euclidean distance of normalized release year and log baseline votes. When fewer than 5 eligible control films exist, I keep all eligible movies.[10] Matching is with replacement, so a control film can serve multiple treated films.

After dropping the few episodes whose event window extends past the end of the ratings data or contains fewer than two days on which I observe both the treated film and at least one of its matched controls, the analysis sample contains 568 treatment events drawn from 560 unique catalog films, together with 1,898 matched control films, for 117,522 film-day observations.[11] Table 1 reports summary statistics for treated and control films in the pre-treatment and post-treatment periods, with the IMDb sample in Panel A and the Letterboxd sample in Panel B. Two features of Panel A

[7] Episodes that cover multiple films, or that do not cover any film (e.g., "mailbag" episodes), are excluded from the analysis.

[8] A second coverage contaminates an event only when it falls within the 30-day event window on either side of the first. No cross-show pair does: the closest, *Rollerball*, is 86 days apart. The 4 events I exclude are multi-part episodes on a single film (*Blank Check*'s two-part *Titanic* and its multi-episode series on each of the three *Star Wars* prequels), whose later parts fall within the window of the first.

[9] IMDb lists up to three genres per film, in alphabetical order, so there is no notion of a film's "primary" genre.

[10] This occurs for 18 of the 568 treatment events; every event retains at least one control, and the mean number of matched controls is 4.9.

[11] The 1,000-vote cutoff excludes 0 otherwise-eligible *Blank Check* films and 6 *HDTGM* films.

Table 1: Summary Statistics

		Baseline	Daily new ratings		Average rating	
	Films	IMDb Votes	Pre	Post	Pre	Post
Panel A: IMDb						
All treated	540	194,791	42.5	48.9	6.34	6.35
Blank Check	336	271,033	65.3	69.8	7.16	7.12
How Did This Get Made?	212	72,783	9.9	15.4	5.09	5.09
All controls	1,831	138,784	28.8	27.2	6.69	6.69
Panel B: Letterboxd						
All treated	261	113,130	2.48	6.12	3.09	2.98
Blank Check	157	170,981	2.54	7.43	3.37	3.32
How Did This Get Made?	104	25,797	2.40	4.15	2.66	2.45
All controls	739	103,855	1.32	1.32	3.34	3.34

Notes: This table reports summary statistics for treated and control films over the 30-day event window. Both panels describe the volume estimation sample. Panel A uses the IMDb sample, and Panel B uses the film-level Letterboxd sample. The pre-treatment period is days −30 to −8 and the post-treatment period is days −7 to 30, matching the treatment onset used in estimation (see Section 4.1). Baseline IMDb Votes is the film's IMDb vote count at the last snapshot before the episode airs, the count used for the vote cutoff and for matching. Average rating is the displayed average on a 1–10 scale in Panel A and the mean of a day's new ratings on a 0.5–5 scale, over days with at least one rated review, in Panel B. A film covered by both shows appears in both shows' rows of Panel A, so those rows sum to more than all treated films.

matter for what follows. First, the two shows cover different types of films: the average *HDTGM* film has 72,783 baseline votes and a 5.09 rating, against 271,033 votes and 7.16 for *Blank Check*. Additionally, the treated films are on average more popular than their control films at baseline, with 194,791 votes and 42.5 new votes per day against 138,784 and 28.8 for the control films. Panel B shows the same pattern on Letterboxd.[12]

3.4 Outcome Variables

I consider two margins of rating activity on each platform. On the *extensive margin*, I capture the effect of coverage on how many people rate using the daily count of new ratings. I estimate the effect on this count by Poisson pseudo-maximum likelihood (PPML) and report each coefficient β as the implied percent change in mean daily new ratings, $100(e^{\beta} - 1)$.[13] On the *intensive margin*,

[12] In the analysis that follows, movie fixed effects absorb any level differences on both platforms.

[13] I use PPML rather than the commonly used $\log(1+\text{new ratings}_{jt})$ approach because the $\log(1+x)$ transformation can produce estimates that are sensitive to the units of x when x is often zero (Chen and Roth, 2024). The number of new ratings is zero for approximately 15.5% of IMDb film-day observations. In Section 5, I report the log OLS estimates for comparison.

I capture how coverage affects what ratings are given. On IMDb, individual votes are unobserved, so the outcome is the displayed average itself, a cumulative stock on a 1–10 scale. The mean of a day's new votes would have to be backed out from day-to-day movements in that average, and the site's 0.1 rounding precision makes that infeasible for catalog films. On Letterboxd, where every new review is observed, the outcome is the daily mean of a film's new ratings. For both Letterboxd outcomes, the new ratings for a given day are those logged on that day. Finally, for the within-reviewer analysis on Letterboxd (see Section 4.5), I use the individual star rating a reviewer assigns to a film.

3.5 Transcript Scoring

To measure what podcasts say about each film, I download podcast episodes from each show's RSS feed and transcribe the audio using OpenAI's Whisper large-v3-turbo model. I score the full episode transcripts using the Claude API.[14] For each episode–film pair, the model rates two dimensions on a $[-1, 1]$ scale: *recommendation valence* (do the hosts encourage watching?) and *quality assessment* (do the hosts consider the film well-made?). I score each transcript eight times and average the draws. The set of assessments for each episode is closely aligned (the median across-draw standard deviation is 0.004 for recommendation valence and 0.007 for the quality assessment), and the average is a continuous measure of each dimension.[15] The predictions developed in Section 2 depend on the quality assessment, which will move the average rating (Prediction 2) and, within reviewer, the individual rating (Prediction 3). While the framework predicts that coverage raises volume whether or not the hosts encourage watching (Prediction 1), recommendation valence captures exactly that encouragement, and I use it to check the prediction in Section 4.4.

Table 2 reports summary statistics for the two transcript dimensions by podcast, and Figure 1 plots each episode's quality assessment against its recommendation valence, by show. The two are positively correlated within each show. *HDTGM* largely occupies the high-recommendation, low-quality quadrant, exactly the pattern the show is predicated on, while *Blank Check* is spread along the diagonal from the low-recommendation, low-quality quadrant to the high-recommendation,

[14] I use Anthropic's `claude-sonnet-4-6` model. See Appendix C for more details on the scoring process.

[15] I separately process the transcripts to determine each episode's *discussion share*, the fraction of the episode devoted to discussing the film, and treat sentiment scores as missing for episodes where it falls below 10%. These episodes stay in the difference-in-differences analysis but drop out of the sentiment-interaction specifications, which require a sentiment score.

Table 2: Transcript Sentiment Score Summary Statistics

Score	Podcast	N	Mean	SD	Min	Max
Panel A: IMDb sample						
Recommendation Valence	All	561	0.459	0.497	-0.86	0.98
	Blank Check	353	0.548	0.463	-0.85	0.98
	HDTGM	208	0.308	0.517	-0.86	0.97
Quality Assessment	All	568	0.120	0.690	-0.95	0.98
	Blank Check	356	0.506	0.523	-0.92	0.98
	HDTGM	212	-0.527	0.384	-0.95	0.88
Panel B: Letterboxd sample						
Recommendation Valence	All	255	0.381	0.513	-0.88	0.98
	Blank Check	155	0.431	0.505	-0.85	0.98
	HDTGM	100	0.304	0.518	-0.88	0.97
Quality Assessment	All	259	-0.007	0.660	-0.95	0.97
	Blank Check	157	0.358	0.565	-0.92	0.97
	HDTGM	102	-0.569	0.307	-0.95	0.61

Notes: This table reports summary statistics for the two transcript-scored sentiment dimensions, each defined on $[-1, 1]$, for the scored episode–film pairs in the IMDb (Panel A) and Letterboxd (Panel B) samples.

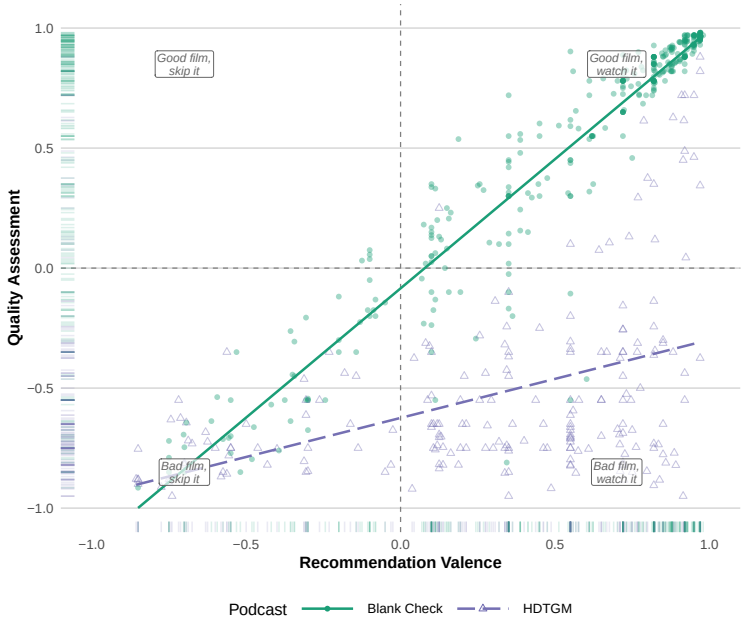


Figure 1: Quality Assessment vs. Recommendation Valence

Notes: This figure plots each episode's transcript-scored quality assessment against its recommendation valence, both on $[-1, 1]$. The lines are linear fits by show, and the marks along the axes show the distribution of each score.

high-quality quadrant, with particular weight in the latter.

4 Empirical Analysis

I estimate the effect of podcast coverage on IMDb and Letterboxd rating activity using a stacked difference-in-differences design. I first describe the estimator and the identifying assumption, which apply throughout. I then estimate the effect of coverage on the volume of new ratings on each platform, after which I consider the effect on the average rating. I next interact treatment with the transcript-scored sentiment measures to link both margins to what the hosts say. Finally, I take advantage of Letterboxd's individual ratings to estimate the within-reviewer specifications, which allow me to separate persuasion from selection.

4.1 Research Design and Identification

With staggered treatment timing across 568 podcast episodes, a standard two-way fixed effects (TWFE) specification uses already-treated units as implicit controls, producing estimates that are variance-weighted averages of potentially heterogeneous treatment effects with weights that can be negative (Goodman-Bacon, 2021). Several heterogeneity-robust estimators address this bias (see, e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021). I use a stacked difference-in-differences design (Cengiz et al., 2019), applying the corrective sample weights of Wing, Freedman, and Hollingsworth (2024).[16]

The stacking approach constructs a separate sub-experiment for each treated film, comparing it only to its matched control films so that the "forbidden comparisons" that bias TWFE are excluded by construction. The corrective weights ensure that treatment and control group trends are weighted equally across sub-experiments. Sub-experiments are then stacked into a single dataset, and the estimating equation is:

$$y_{jgt} = \sum_{k \neq k_{\text{ref}}} \beta_k \cdot \mathbb{1}[\text{relative day} = k] \cdot \text{Treated}_j + \alpha_{jg} + \delta_{tg} + \varepsilon_{jgt}, \tag{4}$$

[16] The computational burden of the Callaway and Sant'Anna (2021) group-time estimator is steep with 568 treatment events occurring on distinct dates; the stacked estimator achieves comparable bias correction at far lower computational cost.

where y_{jgt} is the outcome for movie j in sub-experiment g on day t, Treated_j is an indicator for the covered film, the β_k coefficients trace out the event-study path of the treatment effect relative to day k_{ref} (the omitted reference period), α_{jg} is a movie $\times$ sub-experiment fixed effect, and δ_{tg} is a day $\times$ sub-experiment fixed effect.[17] Standard errors are clustered at the film level, to account for the fact that some control films may appear in multiple sub-experiments. I estimate the design for each margin, the volume of new ratings and their average, on both platforms. Equation (4) traces the dynamic path. For the average effect I replace the day-specific indicators with a single indicator for the post-onset period:

$$y_{jgt} = \beta \cdot \text{Coverage}_{jt} + \alpha_{jg} + \delta_{tg} + \varepsilon_{jgt}, \tag{5}$$

where $\text{Coverage}_{jt} \equiv \text{Treated}_j \cdot \mathbb{1}[t \geq t_0^e]$ turns on for a treated film at the treatment onset t_0^e and remains on through the end of the event window. I set the onset one week before the episode airs, for reasons I describe below.[18]

In order to recover an unbiased estimate of the effect of podcast coverage, I assume that absent podcast coverage, treated and control films would have followed the same trajectory in rating activity and average rating. That is, no new information systematically coincides with podcast episode air dates. The catalog-film setting supports this assumption. These are films released at least 2 years prior with at least 1,000 existing votes, and so their reputations have plausibly settled. Additionally, the fact that I match treated and control films on release year, baseline vote count, and genre helps keep them comparable in rating activity.

Notably, treatment does not necessarily begin when the episode is released. As described in Section 3.1, listeners typically know an episode's film in advance, and many watch the film before they listen. Since some of that viewing will occur in anticipation of the episode's release, the week before air belongs in the treatment window as opposed to the baseline period. In my primary analysis, I therefore set the treatment onset one week before the episode and date the average effect from that point. That is, Coverage_{jt} turns on at relative day -7. The event-study coefficients in

[17] Throughout this analysis, I exclude singleton observations, cases where there is only one observation for a given fixed effect (e.g., a given movie $\times$ sub-experiment or day $\times$ sub-experiment), since they provide no information on the estimated coefficients in the presence of the models' fixed effects.

[18] For comparison, I report two-way fixed effects estimates of the film-level specifications in Appendix B. Estimates from this exercise are consistent with the results presented here.

Equation (4) are normalized to the last pre-onset day, $k_{\text{ref}} = -8$, and I assess pre-trends with a joint F-test on the pre-onset coefficients (β_k for $k < -8$). To confirm that the onset choice does not drive the results, I re-estimate the headline results under alternative onset dates in Appendix A.

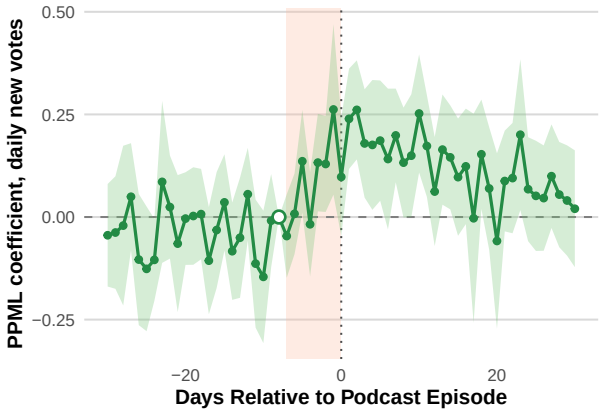
4.2 The Extensive Margin

Podcast coverage raises the volume of new ratings on both platforms, supporting the extensive-margin response of Prediction 1. Table 3 reports the average treatment effects, with IMDb in Panel A and Letterboxd in Panel B. Daily new votes on IMDb rise by 16.4%,[19] and daily new reviews on Letterboxd rise by 127.9%. Panels (a) and (b) of Figure 2 present the volume event studies. On both platforms, volume rises through the week before air. On Letterboxd, volume peaks around the air date, and dissipates over the course of approximately two weeks. The decay is more gradual on IMDb than on Letterboxd, with the latter exhibiting more substantial pre-release viewing. The proportional response is also larger on Letterboxd. The length of the decay is consistent with the asynchronous way in which podcasts are consumed: an episode is heard over the days and weeks after its release, not on a single day. Finally, the rise before air on both platforms is independent corroboration of the pre-air viewing pattern the poll in Section 3.1 describes.[20]

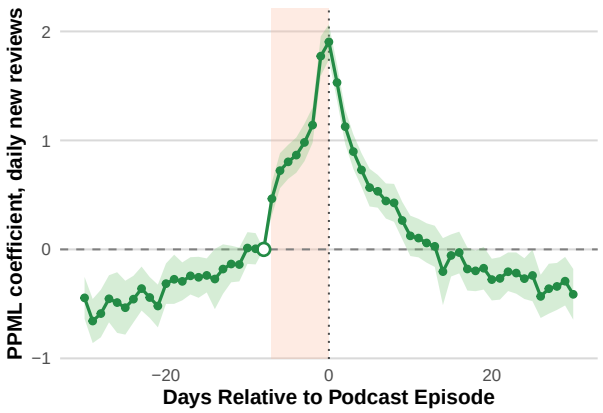
The pre-onset paths in Figure 2 require some discussion, because the joint F-test on the pre-onset coefficients rejects for this outcome on Letterboxd (p <0.001). For IMDb, the joint test over the pre-onset days fails to reject. For Letterboxd, the test restricted to days −14 through −9 does not reject, while the test over all pre-onset days does. Figure 2 shows a smooth drift upward into the event window. One possible explanation is anticipation on a longer horizon, since listeners typically know an episode's film well in advance (Section 3.1); a slow differential drift in treated films' review activity cannot be excluded on this test alone. The Letterboxd estimate is not sensitive to the onset: dating it at day −14 gives an ATT of 133.7%, against 127.9% (Table A.2). As a benchmark for the drift's size, the mean deviation from day −8 over days −30 to −15 is 0.41

[19] The IMDb volume sample in Table 3 is smaller than the rating sample because the first difference of the vote count is undefined on a film's first snapshot and on days when the count falls. The Poisson estimator also drops 19 *Blank Check* episodes due to limited IMDb snapshots.

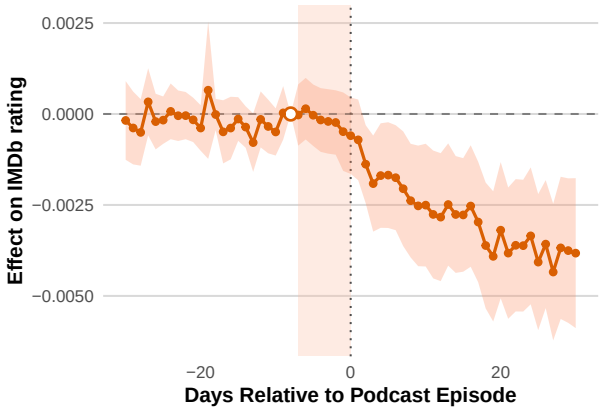
[20] Letterboxd reviewers will occasionally name the podcasts in their written reviews (e.g., "(Blank check podcast homework) My favorite Campion so far, I really enjoyed this film" (holmeskitty, 2022)). In the baseline period, less than one percent of the reviews for covered films name one of the podcasts or hosts. This rate rises in the pre-air anticipatory week, to 1.5 percent for *Blank Check* and 6.7 percent for *HDTGM*. Mirroring the overall volume effect presented here, the likelihood that a review names one of the shows remains elevated through the first post-episode week (2.1 and 6.5 percent, respectively) and then begins to fall back to baseline.



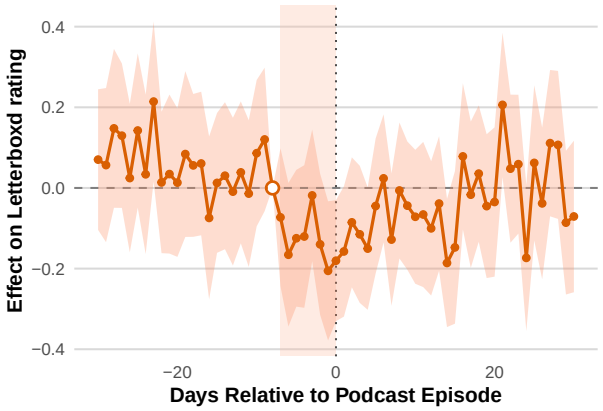
(a) IMDb: Daily New Votes



(b) Letterboxd: Daily New Reviews



(c) IMDb: Average Rating



(d) Letterboxd: Daily Mean of New Ratings

Figure 2: Event Studies: Rating Volume and Average Rating

Notes: This figure plots point estimates and 95% confidence intervals from the stacked DiD specification. Panels (a) and (b) show rating volume: daily new votes on IMDb and daily new reviews on Letterboxd. Volume effects are estimated by Poisson pseudo-maximum likelihood; the implied percent effect is $100(e^{\beta} - 1)$. Panels (c) and (d) show the average rating: the displayed average on IMDb and the daily mean of new reviews' star ratings on Letterboxd. The omitted period is day -8. The shaded region from the treatment onset to the air date is the anticipatory window, and the dotted vertical line marks the air date.

Table 3: Main Results: Effect of Podcast Coverage on Rating Activity

	(1) Rating volume	(2) Average rating
Panel A: IMDb		
Coverage	0.1517***	-0.0021***
	(0.0285)	(0.0007)
Observations	111,568	117,522
Treated films	540	560
Panel B: Letterboxd		
Coverage	0.8237***	-0.1127***
	(0.0588)	(0.0210)
Observations	63,560	31,929
Treated films	261	261

Notes: This table reports stacked DiD average treatment effects. *Coverage* is the treatment indicator of Equation (5), equal to one for a covered film from the treatment onset, one week before the episode air date, onward. Panel A uses the IMDb sample at the film-day level: rating volume is daily new votes, and the average rating is the displayed average on a 1–10 scale. Panel B uses the Letterboxd sample at the film-day level: rating volume is daily new reviews, and the average rating is the daily mean of new reviews' star ratings on a 0.5–5 scale. Volume effects are estimated by Poisson pseudo-maximum likelihood; the implied percent effect is $100(e^{\beta} - 1)$. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

log points, against a post-onset coefficient of 0.824 log points. On IMDb, the rating pre-trend test does not reject. On Letterboxd, it does not reject over either window. I report the joint pre-trend F-tests in Appendix A.

Table 4 reports the IMDb response by film age, baseline popularity, and genre. The increase is largest for the least popular films (64.9% in the lowest baseline-popularity quartile) and smallest for the most popular (9.0% in the highest). This pattern is consistent with coverage bringing a similar number of new raters to every film, since the same number of new votes is a larger share of a smaller film's baseline. The age pattern is monotone: the most recent tercile shows the largest increase (26.8%), against 16.6% for the middle tercile and 11.8% for the oldest. Every genre row shows a statistically significant volume increase, at the 1% level for all but Drama (5%) and Adventure (10%). The response is above the pooled estimate by more than the row's standard error for Action, Crime, and Thriller films, and below it by more than that for Drama and Adventure films. On the rating, Drama and Thriller films show the deepest declines. The rating decline is largest in the lowest-popularity quartile, and it is significant at the 10% level in each of the three age groups.

Table 4: Heterogeneity by Film Characteristics (IMDb)

		Rating volume		Average rating	
Dimension	Group	Films	ATT	Films	ATT
Age	Recent (<10y)	94	0.2372**	98	-0.0029**
			(0.0925)		(0.0014)
Age	Classic (10-25y)	195	0.1532***	202	-0.0024*
			(0.0440)		(0.0013)
Age	Vintage (>25y)	253	0.1119***	262	-0.0017*
			(0.0308)		(0.0009)
Popularity	Q1 (Low)	126	0.5002***	132	-0.0056***
			(0.0590)		(0.0019)
Popularity	Q2	125	0.2698***	131	-0.0000
			(0.0610)		(0.0007)
Popularity	Q3	130	0.2759***	131	-0.0001
			(0.0867)		(0.0005)
Popularity	Q4 (High)	132	0.0861***	132	-0.0015*
			(0.0298)		(0.0008)
Genre	Drama	232	0.0627**	248	-0.0036***
			(0.0252)		(0.0012)
Genre	Action	186	0.2888***	188	-0.0016
			(0.0581)		(0.0010)
Genre	Comedy	170	0.1897***	177	-0.0028**
			(0.0663)		(0.0012)
Genre	Adventure	159	0.0723*	161	-0.0013
			(0.0384)		(0.0011)
Genre	Crime	107	0.2079***	110	-0.0011
			(0.0549)		(0.0014)
Genre	Thriller	98	0.2569***	101	-0.0038**
			(0.0620)		(0.0016)

Notes: This table reports stacked DiD average treatment effects on IMDb by film age tercile, baseline-popularity quartile, and genre. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes; the implied percent effect is $100(e^{\beta} - 1)$. Each Films column counts the treated films in the sample of the regression to its right; a film covered by both shows is counted at each coverage. Age is measured at the air date, and baseline popularity is the film's mean vote count over days -30 to -8. Genre rows are not exclusive: a film appears in the row of every genre it lists, and the table shows every genre listed for at least 100 covered films. Each event's control films enter its row. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4.3 The Intensive Margin

I now estimate the effect of coverage on the average rating, the intensive-margin response of Prediction 2, and find that it falls on both platforms (column (2) of Table 3). On IMDb, the effect on the average rating is -0.0021 points on the 1–10 scale, a small but statistically significant decline. The size of the effect is muted by construction. Recall that the IMDb outcome is a cumulative average, constructed from tens of thousands of prior votes, so a couple of weeks of podcast-driven votes can only move it by so much. Moreover, because I observe the average only after the site rounds it to a tenth, much of that movement is undetectable in my data. On Letterboxd, the effect on the daily mean of that day's new ratings is -0.113 stars on the 0.5–5 scale.[21]

Panels (a) and (b) of Figure 2 show daily votes begin rising a week before air, as listeners watch in anticipation of the episode. Over that anticipatory period, panels (c) and (d) show the average rating drifts slightly below baseline, although the coefficients remain indistinguishable from zero. After the episode is released, the decline on IMDb becomes statistically significant and roughly triples in depth over the following week. The Letterboxd daily average deepens as well, although its path is noisier. These findings are consistent with both persuasion and selection: a somewhat selected crowd arrives ahead of the episode's release, and the ratings decline accelerates following the hosts' discussion. Of course, timing alone cannot separate the two, which I return to below.

To better understand where in the rating distribution the decline originates, I decompose the Letterboxd review flow into daily counts by star tier—low (≤ 2.5), mid (3–3.5), and high (≥ 4)—and estimate Equation (5) on each tier's count. I present the results of this exercise in Figure 3. Coverage increases the number of reviews for all three tiers, but low-star reviews rise by 52% more than high-star reviews at the release date peak. As a result, the average rating falls because low ratings are added faster than high ones. The fact that the high tier rises alongside the others indicates that coverage mobilizes praise alongside pans, consistent with Prediction 1. Additionally, the uneven rise is a further rejection of the pure attention mechanism, under which every tier would scale by the same factor. What this decomposition cannot distinguish is whether the decline in ratings is due to harsher reviewers arriving or due to the typical reviewers rating lower than they

[21] An episode also tells a listener a good deal about a film, so it may steer away some of those who would dislike it. On its own, that sorting would raise the average rating among those who watch, so it works against the decline I estimate.

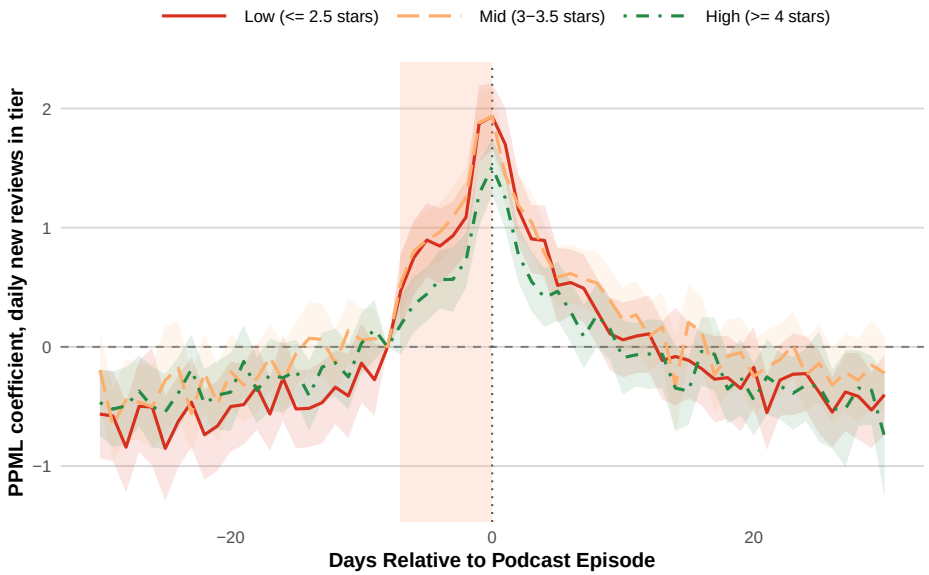


Figure 3: Composition of the Letterboxd Review Flow by Star Tier

Notes: This figure plots stacked DiD event studies for the daily review count within three star tiers (low ≤ 2.5, mid 3–3.5, high ≥ 4) on Letterboxd; a tier's count is zero on days with no review in it. Volume effects are estimated by Poisson pseudo-maximum likelihood; the implied percent effect is $100(e^{\beta} - 1)$. Each tier is estimated on its own sample, which drops the cells in which that tier's count is always zero. Line types distinguish the tiers, and the shaded bands are 95% confidence intervals. The shaded region from the treatment onset to the air date is the anticipatory window, and the dotted vertical line marks the air date.

otherwise would. Separating the two requires holding the reviewer fixed, which I do in Section 4.5.

4.4 The Content Gradient

I next interact the treatment indicator with transcript-derived sentiment scores to understand how the treatment effect varies with what the hosts say (Prediction 2):

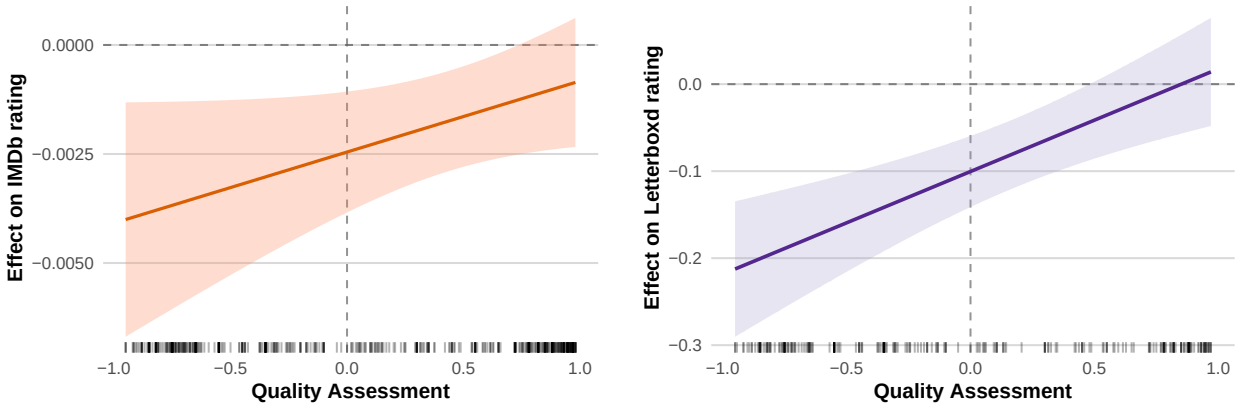
$$y_{jgt} = \underbrace{(\beta_1 + \beta_2 \cdot s_e)}_{\tau(s_e)} \cdot \text{Coverage}_{jt} + \alpha_{jg} + \delta_{tg} + \varepsilon_{jgt}, \tag{6}$$

where s_e is one of the two sentiment scores for the treating episode, recommendation valence or quality assessment, each entering a separate regression, $\tau(s_e)$ is the treatment-effect function, β_1 is the post-air effect at $s_e = 0$, and β_2 is the between-film slope. A positive β_2 indicates that covered films the hosts judge lower end up rated lower. Importantly, this is a between-film comparison—covered films the hosts pan against covered films they praise. Since β_2 captures the effect of what the hosts say, its pre-treatment baseline must end before the podcast begins to draw attention to the film, i.e., prior to the start of the anticipatory week. Additionally, its treated period cannot begin until the episode is released to listeners. Thus, the content-gradient specifications date coverage from the air date, when the episode's discussion becomes available, and drop each covered film's observations in the anticipatory week before air, so the baseline is days -30 to -8.

I find that the intensive-margin decline is larger the more negatively the hosts assess the film. Figure 4 plots the treatment-effect function τ(quality assessment) on each platform. The slope is positive on both platforms, though larger on Letterboxd.[22] The more negatively the hosts judge a film, the further its rating falls. The estimates reject pure attention, under which raters drawn from a film's existing assessment distribution would produce no slope at all (Prediction 2).

In Appendix A.2, I report the corresponding gradient for the extensive margin, interacting coverage with recommendation valence (Table A.3). Prediction 1 concerns the sign of the volume effect across the range of the hosts' encouragement, not its slope. On IMDb, I fail to find evidence that the volume response varies with recommendation valence. On Letterboxd, the slope is negative at the 10% level, so the volume response is somewhat larger for the films the hosts steer listeners

[22] This pattern is not driven by the linear assumption: estimating the effect separately by tercile of the quality score gives estimates close to the fitted line for both platforms (see Figure A.1).



(a) IMDb

(b) Letterboxd

Figure 4: Treatment Effect by Quality Assessment

Notes: This figure plots the linear interaction estimate of $\tau(\text{quality assessment}) = \beta_1 + \beta_2 \cdot \text{quality}$ on the average rating, on IMDb (panel (a)) and Letterboxd (panel (b)). Shaded areas are 95% confidence bands, and the marks along the horizontal axis show the covered films' scores. A positive slope means films the hosts judge lower are rated lower after coverage.

away from. On both platforms, the implied effect remains positive across the observed range of scores, so coverage raises volume for praised and panned films alike.

Table 5 estimates Equation (5) separately for each show, with IMDb in Panel A and Letterboxd in Panel B. On IMDb, both raise volume, with a larger effect for *HDTGM*. On the intensive margin, the *HDTGM* rating ATT is -0.0024 points against -0.0017 for *Blank Check*, though I fail to find evidence that the two effects are meaningfully different. Panel B reports the same comparison on Letterboxd. Both shows again raise volume there, this time with a larger effect for *Blank Check*. The *HDTGM* rating ATT is again larger in magnitude than the *Blank Check* ATT, though I am unable to conclude that the two intensive-margin effects are statistically different. The *HDTGM* result is the sharpest test of Prediction 1: its hosts assess the films they cover as poor on average, and coverage raises volume on both platforms regardless.

Importantly, the evidence presented so far comes from between-film comparisons, so it shows that the intensive-margin response tracks the content of coverage, but it cannot show that content actually changed anyone's mind. If a show's audience is relatively stable, so the same listeners watch and rate whichever film their show covers, then the results presented here are likely driven by persuasion. If instead each episode draws its own crowd, then I am again unable to separate

Table 5: *HDTGM* vs. *Blank Check*

	HDTGM	Blank Check	Difference
Panel A: IMDb			
ATT Estimates			
Rating volume (daily new votes)	0.3807***	0.0723***	0.3084***
	(0.0711)	(0.0248)	(0.0753)
Average rating	-0.0024**	-0.0017***	-0.0008
	(0.0010)	(0.0006)	(0.0012)
Mean Sentiment Scores			
Recommendation Valence	0.31	0.55	-0.24
Quality Assessment	-0.53	0.51	-1.03
Sample			
Treated films (volume)	212	336	
Treated films (rating)	212	356	
Panel B: Letterboxd			
ATT Estimates			
Rating volume (daily new reviews)	0.4974***	1.0482***	-0.5508***
	(0.0895)	(0.0657)	(0.1110)
Average rating	-0.1491***	-0.0892***	-0.0599
	(0.0371)	(0.0250)	(0.0447)
Mean Sentiment Scores			
Recommendation Valence	0.30	0.43	-0.13
Quality Assessment	-0.57	0.36	-0.93
Sample			
Treated films (volume)	104	157	
Treated films (rating)	104	157	

Notes: This table reports stacked DiD estimates separately for *HDTGM* and *Blank Check* episodes, with IMDb in Panel A and Letterboxd in Panel B. The Difference column is *HDTGM* minus *Blank Check*. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes (reviews); the implied percent effect is $100(e^{\beta} - 1)$. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the effects of self-selection and persuasion.[23]

4.5 The Within-Reviewer Response

In order to disentangle the persuasion and selection mechanisms, I leverage the individual review-level Letterboxd ratings, to estimate the effect of the hosts' assessment on a reviewer's rating while holding the reviewer fixed. Let r_{ijgt} be the star rating that reviewer i assigns film j in sub-experiment g on day t, and let s_e be the quality assessment of the episode e covering film j. The pooled specification omits the reviewer:

$$r_{ijgt} = \beta \cdot \text{Coverage}_{jt} + \alpha_{jg} + \delta_{tg} + \varepsilon_{ijgt}. \tag{7}$$

I next add a reviewer fixed effect, μ_i, which differences out anything fixed about the reviewer, such as their critical disposition. With this fixed effect in place, β is identified only from reviewers who are observed both at baseline (i.e., on a control film, or on a covered film before onset) and on a covered film after coverage:

$$r_{ijgt} = \beta \cdot \text{Coverage}_{jt} + \alpha_{jg} + \delta_{tg} + \mu_i + \varepsilon_{ijgt}. \tag{8}$$

My preferred specification interacts the treatment with episode sentiment while holding the reviewer fixed:

$$r_{ijgt} = \underbrace{(\beta + \theta \cdot s_e)}_{\tau(s_e)} \cdot \text{Coverage}_{jt} + \alpha_{jg} + \delta_{tg} + \mu_i + \varepsilon_{ijgt}. \tag{9}$$

The coefficient of interest is θ, the within-reviewer slope of Prediction 3, which captures how the same reviewer's rating of a covered film moves with how the hosts assessed it. As in Section 4.4, the content-gradient specification dates coverage from the air date and drops the covered film's ratings dated in the week before air, so the baseline is days -30 to -8.[24] A reviewer who rates a covered film after release but has not yet listened to the episode is untreated by its content, so such ratings bias my estimate of θ toward zero, making it an understatement of the effect of the episode's

[23] Unfortunately, audience composition data is not available.

[24] On Letterboxd the week before air carries a second problem: the platform dates a rating by the day the reviewer watched the film, not the day the review was posted, so a rating dated in that week may have been written after the episode aired.

Table 6: Within-Reviewer Estimator: Nested Specifications (Letterboxd)

	(1)	(2)	(3)	(4)	(5)	(6)
Coverage (β)	-0.1138***	0.0002	-0.1060***	-0.0222		-0.0180
	(0.0161)	(0.0236)	(0.0167)	(0.0306)		(0.0305)
Coverage × quality (θ)			0.0968***	0.1294***	0.0985*	0.0816*
			(0.0254)	(0.0438)	(0.0537)	(0.0469)
Likely listener × quality						0.1843***
						(0.0586)
Reviewer fixed effect	No	Yes	No	Yes	Yes	Yes
Coverage × show	No	No	No	No	Yes	No
Reviews	128,495	68,690	109,750	51,236	51,236	51,236
Reviewers	72,076	14,583	68,970	13,045	13,045	13,045
Identifying reviewers	–	6,341	–	7,387	7,387	7,387

Notes: This table reports review-level estimates on Letterboxd. Column (3) is the between-film specification, the pooled regression with the sentiment interaction, and column (4) is its within-reviewer counterpart; Figure 5 plots the two. Coverage turns on at day −7 in columns (1) and (2) and at the air date in columns (3) through (6), which drop the anticipatory week and exclude two covered films with no transcript score. Column (6) adds the interaction of the quality assessment with an indicator for likely listeners, reviewers who rate a covered film on or after day −7. Identifying reviewers are those whose reviews identify the treatment coefficients: reviewers observed both at baseline and on a covered film after coverage in column (2), and reviewers who rate a covered film after coverage and at least one other film in columns (4) through (6). Singleton observations are excluded. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

content on listeners. Because the manner in which the two shows approach and discuss movies is notably different, θ could reflect a difference between the two shows rather than the relationship between episode content and rating behavior. I therefore also estimate a version of Equation (9) that allows the coverage effect to differ by show, so θ is identified only from within-show differences in quality assessments. I estimate these specifications following the same approach as in every other regression in this paper.

I present the results in Table 6. Column (1) is the review-level counterpart of the Letterboxd rating effect in Table 3, and the two estimates match. Column (2) adds the reviewer fixed effect, and the coverage effect disappears. That is, the same reviewer does not, on average, rate a covered film lower after coverage. The difference between columns (1) and (2) demonstrates the selection channel. The decline estimated in column (1) reflects *who* rates after coverage, not a uniform shift in how a given reviewer rates. Whether the content of an episode moves a given reviewer is a separate question, which I turn to next.[25]

[25]I exclude reviewers who only appear once, since their single review will be entirely captured by their fixed effect,

Column (3) interacts coverage with the episode's quality assessment without the reviewer fixed effect, which is the between-film gradient of Section 4.4 estimated at the review level. Column (4) adds the reviewer fixed effect. The within-reviewer sentiment slope is $\theta = 0.129$, which indicates that the same reviewer rates a more-panned film lower and a more-praised one higher. Moving the hosts' assessment across its observed range moves the same reviewer's rating by 0.23 stars, a little over a quarter of the typical spread in a single reviewer's ratings (0.83 stars). I present the result of the within-show specification in column (5). The within-show slope is slightly smaller than the pooled slope in column (4), but remains positive and statistically significant. Taken together, these results support the hypothesis that the content of the podcast episodes—how the hosts and guests discuss and evaluate the movie—meaningfully affects the audience's perception of the movie.

In Figure 5, I plot the fitted coverage effect as a function of the episode's quality assessment, estimated between films and within reviewers. The between-film slope (column (3) of Table 6, 0.097) is the total content gradient, and the within-reviewer slope (column (4), 0.129) is what remains once the reviewer is held fixed. If the between-film gradient were driven by reviewers sorting across films by how harshly or generously they rate, the slope would shrink once the reviewer fixed effect is added. If anything, it is larger. Of course, the reviewer fixed effect removes a reviewer's average rating level, and with it any sorting of reviewers across films by that level, such as the tendency Bondi, Rossi, and Stevens (2025) document for the most experienced raters on IMDb and MovieLens to choose higher-quality films and to rate nearly every film more harshly than novices do. It does not, however, remove within-person "occasion-selection"—the reviewer's choice of when to watch and rate a film—which I discuss in Section 6.

While the reviewer fixed effect removes level differences between the podcast-listening audience and the general public, it does not address the possibility that listeners' tastes track the hosts' more closely than other reviewers' tastes do. To the extent that they do, listeners' ratings of covered films would correlate with the hosts' assessment even before the episode is released, resulting in a column (4) estimate that conflates this selection with persuasion. To address this concern, I identify reviewers who are likely podcast listeners ("likely listeners") as those who rate at least one

and as a result they provide no information on the estimated coefficients. I similarly exclude a small number of reviews that are the only observation for a given film or day within a sub-experiment. This leaves 51,236 reviews across 13,045 reviewers in columns (4) through (6). The slope is identified by the 7,387 reviewers who rate a covered film after coverage and at least one other film in the sample; they write 37,146 reviews.

covered film from the week before its episode airs onward (i.e., during the anticipatory week or the post-release period). By construction, every post-release rating of a covered film comes from a likely listener, so θ is identified only from their ratings. To capture the possibility that likely listeners' tastes are positively correlated with the hosts', I add the interaction of the hosts' quality assessment with an indicator for likely listeners to Equation (9), applied to every rating of a covered film before and after release:

$$r_{ijgt} = (\beta + \theta \cdot s_e) \cdot \text{Coverage}_{jt} + \gamma \cdot \text{Listener}_i \cdot \text{Treated}_j \cdot s_e + \alpha_{jg} + \delta_{tg} + \mu_i + \varepsilon_{ijgt}. \tag{10}$$

The coefficient γ measures how much more closely likely listeners' ratings of a covered film track the hosts' assessment than other reviewers' ratings do before the episode airs. As a result, θ now measures the *change* in likely listeners' slope once the episode is released. Since the episode's content cannot persuade a listener before it is released, γ reflects the tastes of the audience (i.e., how closely listeners' tastes track the hosts'), and θ reflects the persuasive effect of the episode's discussion.

I report the estimates of Equation (10) in column (6) of Table 6. I find that likely listeners' ratings track the hosts' assessment more closely than other reviewers' ratings before coverage. This is consistent with listeners selecting into a show's audience not only on how harshly or generously they rate, which the reviewer fixed effect absorbs, but also on how closely their tastes track the hosts'. Once the episode airs, however, likely listeners' slope rises by 44%. That is, the hosts' specific discussion of a film during the episode moves the ratings of their audience, even though that audience's tastes already track the hosts' closely.

Altogether, the results presented in this section support the three predictions developed in Section 2. Media coverage raises the volume of new ratings on both platforms, regardless of whether the hosts praise or pan the film (Prediction 1). The average rating falls, and that fall is steeper the lower the hosts' assessment of the film (Prediction 2). Finally, holding the reviewer's idiosyncratic preferences fixed, the same person rates a covered film in the direction of that assessment (Prediction 3). That is, the hosts' discussion of a film changes how a given listener rates it: coverage alters how people perceive the films they watch, and therefore how they rate them.

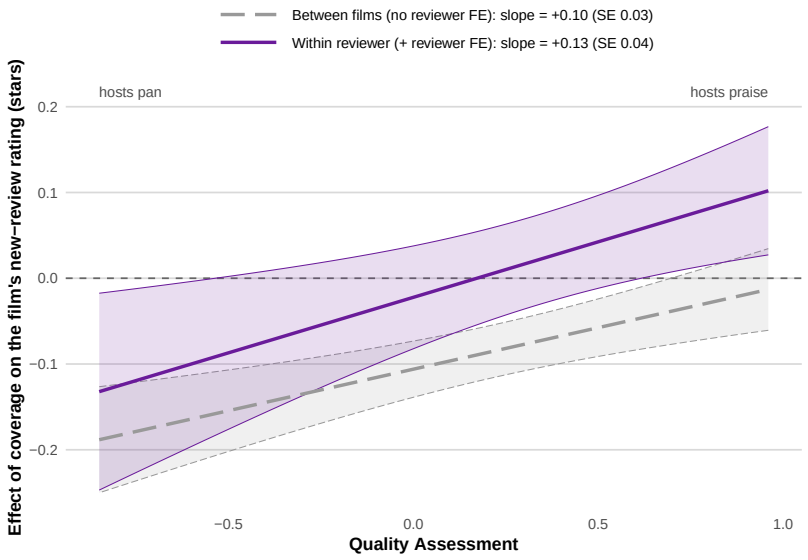


Figure 5: The Coverage Effect on Ratings, Between Films and Within Reviewers (Letterboxd)

Notes: This figure plots the effect of podcast coverage on a reviewer's star rating on Letterboxd as a function of the episode's quality assessment, estimated between films and within reviewers (columns (3) and (4) of Table 6). Shaded areas are 95% confidence bands.

5 Robustness

In this section, I summarize a set of robustness checks of the estimates in Section 4. First, I conduct a placebo test that moves the air dates, then I run a check that restricts the IMDb sample to the period with daily data coverage. Finally, I re-estimate the volume effect by OLS on the log of one plus daily new ratings, as a check on functional form. The analyses summarized in this section are described in detail in Appendix A, which also reports the pre-trend and onset-sensitivity tables discussed in Section 4.

5.1 Placebo Tests

One possible concern is that podcast episodes are released when attention to films like the treated one is rising for other reasons—a seasonal interest, a streaming service adding a set of films from the same genre or era, a sequel reviving interest in a category of films, etc. If that is the case, part of the estimated response would reflect the concurrent changes rather than the podcast coverage. The matched control films address this concern. They were released within 3 years of the covered film, have a similar baseline vote count, and share at least one of its genres, so a wave that lifts

films of that kind should lift them too, and the matched comparison differences it out. Of course, this only works if the covered films and their controls actually move together when no episode airs. I check that directly by shifting the podcast release dates earlier by 60 days, so that the event window closes a month before the episode airs. I then re-estimate the stacked DiD specification around these false dates. I report the results in Table A.4 in Appendix A.3. The placebo effect on the volume of new votes is 0.0053 (SE 0.0240), an implied change of 0.5% against the 16.4% I estimate at the actual air dates, and it is indistinguishable from zero. The placebo effect on the average rating is 0.0000 (SE 0.0005). This exercise provides evidence that the covered films and their controls do not drift apart on their own, and that the response in Section 4 is tied to the air dates themselves. That said, a placebo test cannot speak to news about the covered film itself that arrives alongside the episode, for instance if the hosts choose a film because a sequel is about to be released. The way the two shows choose films, described in Section 3.1, limits the scope for this concern. *Blank Check* covers each director's films in release order, so a film's air date is set by its place in the miniseries rather than by news about the film. *HDTGM* chooses its films one at a time, which leaves more room for this kind of timing, but such coincidences are rare in both shows: only 16 of the 568 events have a sequel, remake, or other film in the same franchise released within 30 days of the episode.

5.2 Data Coverage

As described in Section 3.2, the IMDb ratings record is sparse before January 21, 2023 and daily thereafter. In light of the fact that the flow outcome across a gap between snapshots is an average rather than a single day's count, and because the sparse events contribute few observations to the event-study bins, I re-estimate the main IMDb specifications on the 319 podcast episodes that air on or after January 21, 2023. I report the results of this in Table A.5 in Appendix A.4. Coverage raises the volume of new votes by 19.3% in the daily-coverage sample, against 16.4% in the full sample, and the effect on the average rating is -0.0020 (SE 0.0007) against -0.0021 (SE 0.0007).

5.3 Functional Form of Rating Volume

As noted in Section 3, I estimate the volume effect by PPML because the $\log(1+x)$ transformation can produce estimates that are sensitive to the units of the outcome when it is often zero (Chen

and Roth, 2024). As a check on functional form and to relate my estimates to a commonly used approach, I re-estimate the effect by OLS with the log of one plus daily new ratings as the outcome. I report the results of this in Appendix A.5 side-by-side with the PPML estimates. The two estimators agree in sign and significance on both platforms. On IMDb, they also agree in magnitude, with a log OLS coefficient of 0.1768 log points against a PPML coefficient of 0.1517. On Letterboxd, they do not. There, the log OLS coefficient of 0.452 log points is a little over half the PPML coefficient of 0.824. This difference reflects exactly the concerns raised by Chen and Roth (2024): daily new reviews are zero for 38.7% of Letterboxd film-days, against 15.5% on IMDb, and as a result the log transformation compresses its proportional change.

6 Concluding Discussion

This paper provides evidence on how media coverage shapes the ratings that online platforms report. Studying two popular movie podcasts, I find that coverage raises the volume of new ratings by 16.4% on IMDb and 127.9% on Letterboxd and lowers the films' average ratings on both platforms. Holding the reviewer fixed, I find that the content of coverage affects ratings.

At the film level, persuasion and selection are observationally equivalent (cf. DellaVigna and Gentzkow, 2010). I find that when the average rating falls, the fall deepens with how negatively the hosts assess the film, and the added reviews arrive disproportionately in the lower star tiers. While these patterns together reject pure attention, under which raters drawn from a film's existing assessment distribution would produce no gradient at all, both are consistent with the hosts changing who rates and changing the minds of those who rate. Using individual review-level data, I find that the content of the hosts' discussions directly affects ratings—moving the hosts' assessment across its observed range shifts the same reviewer's rating by 0.23 stars, a little over a quarter of the 0.83-star spread in a typical reviewer's own ratings. This slope is, if anything, larger than its between-film counterpart.

Several limitations qualify these findings. First, while holding the reviewer fixed controls for time-invariant user characteristics, it cannot account for the reviewer's choice of when to watch and rate a film. The threat of this "occasion-selection" concern is likely to be limited. A general tendency to rate more harshly after listening would shift the level of a reviewer's ratings but not

the slope. To affect the estimated slope, this concern would have to mirror the podcast's quality assessment film by film.

The Letterboxd ratings I analyze are also selected. The ratings I observe are those that accompany written reviews. To the extent that those reviewers hold stronger opinions than users who rate but do not write a review, the within-reviewer response will likely understate what the full population of raters would show. And while the podcast hosts' sentiment is not randomly assigned, the within-reviewer design differences out film-level and reviewer-level confounds, leaving occasion-selection as the residual concern I discuss above. Finally, the estimates capture the direct effect of podcast coverage. Any social multiplier à la Moretti (2011)—say, a listener's activity prompting friends to watch and rate the same film—would make the estimates I present here an understatement of the total effect of podcast coverage.

These findings are not specific to movie podcasts. The ratings of many products sit within a wider ecosystem of attention intermediaries, including critics, vloggers, social feeds, and, of course, podcasters. Coverage from any of them can both draw new consumers to a product and affect their judgment of it. The cumulative effect of many such treatments determines how independent a displayed average is of the coverage the product receives. The approach developed here, pairing attention events with transcript-scored measures of their content, provides a template for future work to measure it in other markets.[26]

This work recasts a common platform input that is usually treated as fixed. Consumers, researchers, and ranking algorithms often read a displayed average as a stable summary of quality (Hu, Pavlou, and Zhang, 2017; Luca, 2016), particularly for longstanding products. But both the set of people who rate and the ratings they give respond to the media environment around the platform. Moreover, the exposure is uneven: the volume response is largest for films in the lowest baseline-popularity quartile. This puts products with thin rating histories at the highest risk of being influenced by media coverage. Platforms could respond, discounting rating surges that coincide with identifiable attention events or displaying the uncertainty behind the number, though whether such adjustments would improve the information a rating conveys is an open question. Indeed, to the extent that coverage genuinely persuades listeners, the ratings that result are an

[26] See, for example, Ash et al. (2026), who take a related approach, pairing shocks to listenership of *The Joe Rogan Experience* with model-coded transcripts of its episodes.

accurate reflection of their perception of the film, and it is not clear that a platform would want to discount them.

Ultimately, the reputations that platforms report are shaped by the media environment around them. This analysis has shown that a product's average rating reflects both who was motivated to rate and what they were persuaded to say. Treating that rating as a fixed, exogenous signal of quality ignores an input that responds, measurably, to as little as a single podcast episode.

References

Anderson, Michael and Jeremy Magruder. 2012. "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *Economic Journal* 122 (563):957–989.

Aridor, Guy, Duarte Gonçalves, Daniel Kluver, Ruoyan Kong, and Joseph Konstan. 2026. "The Informational Role of Online Recommendations: Evidence from a Field Experiment." Discussion Paper 21584, CEPR.

Ash, Elliott, Sergio Galletta, Federico Masera, and Lehan Zhang. 2026. "The Joe Rogan Effect: Politicized Podcasts and the Youth Gender Voting Gap." Working Paper, ETH Zürich.

Berger, Jonah, Alan T. Sorensen, and Scott J. Rasmussen. 2010. "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales." *Marketing Science* 29 (5):815–827.

Bondi, Tommaso, Michelangelo Rossi, and Ryan L. Stevens. 2025. "The Good, the Bad and the Picky: Consumer Heterogeneity and the Reversal of Product Ratings." *Management Science* 71 (8):7200–7222.

Brandes, Leif, Stephan Nüesch, and Egon Franck. 2016. "Death-Related Publicity as Informational Advertising: Evidence from the Music Industry." *Marketing Letters* 27 (1):143–157.

Callaway, Brantly and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2):200–230.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *Quarterly Journal of Economics* 134 (3):1405–1454.

Chen, Jiafeng and Jonathan Roth. 2024. "Logs with Zeros? Some Problems and Solutions." *Quarterly Journal of Economics* 139 (2):891–936.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review* 100 (4):1358–1398.

Chevalier, Judith A. and Dina Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research* 43 (3):345–354.

Cosley, Dan, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. "Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 585–592.

Dellarocas, Chrysanthos. 2003. "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms." *Management Science* 49 (10):1407–1424.

Dellarocas, Chrysanthos and Charles A. Wood. 2008. "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias." *Management Science* 54 (3):460–476.

DellaVigna, Stefano and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics* 2 (1):643–669.

Eliashberg, Jehoshua and Steven M. Shugan. 1997. "Film Critics: Influencers or Predictors?" *Journal of Marketing* 61 (2):68–78.

Filippas, Apostolos, John J. Horton, and Joseph M. Golden. 2022. "Reputation Inflation." *Marketing Science* 41 (4):733–745.

Garthwaite, Craig L. 2014. "Demand Spillovers, Combative Advertising, and Celebrity Endorsements." *American Economic Journal: Applied Economics* 6 (2):76–104.

Gilchrist, Duncan Sheppard and Emily Glassberg Sands. 2016. "Something to Talk About: Social Spillovers in Movie Consumption." *Journal of Political Economy* 124 (5):1339–1382.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2):254–277.

He, Sherry, Brett Hollenbeck, and Davide Proserpio. 2022. "The Market for Fake Reviews." *Marketing Science* 41 (5):896–921.

Historical IMDb Ratings. 2026. "Historical IMDb Ratings API." https://api.historical-imdb-ratings.com/swagger/index.html. Accessed April 13, 2026.

holmeskitty. 2022. "Review of *Bright Star*." Letterboxd, https://letterboxd.com/holmeskitty/film/bright-star/. Review dated February 26, 2022.

Horton, Donald and R. Richard Wohl. 1956. "Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance." *Psychiatry* 19 (3):215–229.

Hu, Nan, Paul A. Pavlou, and Jie Zhang. 2017. "On Self-Selection Biases in Online Product Reviews." *MIS Quarterly* 41 (2):449–471.

IMDb. 2026a. "IMDb Non-Commercial Datasets." https://datasets.imdbws.com. Accessed April 13, 2026.

———. 2026b. "Our Family of Companies." https://help.imdb.com/article/imdbpro/membership-benefits/our-family-of-companies/GKKSCH4H4KARV2BA. Accessed October 5, 2026.

———. 2026c. "Ratings FAQ." https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV. Accessed April 13, 2026.

Kovács, Balázs and Amanda J. Sharkey. 2014. "The Paradox of Publicity: How Awards Can Negatively Affect the Evaluation of Quality." *Administrative Science Quarterly* 59 (1):1–33.

Leyden, Benjamin T. 2025. "Platform Design and Innovation Incentives: Evidence from the Product Rating System on Apple's App Store." *International Journal of Industrial Organization* 99:103133.

Li, Xingyi, Yiting Deng, Puneet Manchanda, and Bert De Reyck. 2026. "Can Lower(ed) Expert Opinions Lead to Better Consumer Ratings?: The Case of Michelin Stars." *Management Science* 72 (3):2427–2450.

Li, Xinxin and Lorin M. Hitt. 2008. "Self-Selection and Information Role of Online Product Reviews." *Information Systems Research* 19 (4):456–474.

Liu, Yong. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue." *Journal of Marketing* 70 (3):74–89.

Luca, Michael. 2016. "Reviews, Reputation, and Revenue: The Case of Yelp.com." Working Paper 12-016, Harvard Business School.

Luca, Michael and Georgios Zervas. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science* 62 (12):3412–3427.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review* 104 (8):2421–2455.

Moe, Wendy W. and David A. Schweidel. 2012. "Online Product Opinions: Incidence, Evaluation, and Evolution." *Marketing Science* 31 (3):372–386.

Moe, Wendy W. and Michael Trusov. 2011. "The Value of Social Dynamics in Online Product Ratings Forums." *Journal of Marketing Research* 48 (3):444–456.

Moretti, Enrico. 2011. "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales." *Review of Economic Studies* 78 (1):356–393.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor. 2013. "Social Influence Bias: A Randomized Experiment." *Science* 341 (6146):647–651.

Nosko, Chris and Steven Tadelis. 2026. "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment." *Quantitative Marketing and Economics* 24 (1):9.

Reimers, Imke and Joel Waldfogel. 2021. "Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings." *American Economic Review* 111 (6):1944–1971.

Reinstein, David A. and Christopher M. Snyder. 2005. "The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics." *Journal of Industrial Economics* 53 (1):27–51.

Rossi, Michelangelo and Felix Schleef. 2026. "Quality Disclosures and Disappointment: Evidence from the Academy Nominations." *Management Science*, Articles in Advance.

Scheer, Paul, June Diane Raphael, and Jason Mantzoukas. 2012. "Speed 2: Cruise Control." *How Did This Get Made?* podcast, Earwolf. Episode 37, with Scott Aukerman.

———. 2017. "The Garbage Pail Kids Movie: LIVE!" *How Did This Get Made?* podcast, Earwolf. Episode 169, with Jon Lovett. Recorded live at Largo at the Coronet, Los Angeles.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl. 2020. "The Polarity of Online Reviews: Prevalence, Drivers and Implications." *Journal of Marketing Research* 57 (5):853–877.

Seiler, Stephan, Song Yao, and Wenbo Wang. 2017. "Does Online Word of Mouth Increase Demand? (And How?) Evidence from a Natural Experiment." *Marketing Science* 36 (6):838–861.

Sun, Liyang and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2):175–199.

Tadelis, Steven. 2016. "Reputation and Feedback Systems in Online Platform Markets." *Annual Review of Economics* 8 (1):321–340.

Troile. 2023. "Do you listen to episodes without having seen the movie?" https://www.reddit.com/r/blankies/comments/10us7k4/. Reddit poll, r/blankies (*Blank Check*) subreddit. Accessed May 29, 2026.

Wing, Coady, Seth M. Freedman, and Alex Hollingsworth. 2024. "Stacked Difference-in-Differences." Working Paper 32054, National Bureau of Economic Research.

Zuloaga Cosme, Francisco, Paul Parker, and Jonathan Beck. 2025. "Weaponized Ratings: The Revenue Impact of Review Bombing in the Entertainment Market." Working Paper 5336870, SSRN.

Table A.1: Pre-Trend Tests (Stacked DiD)

	Rating volume		Average rating	
Pre-onset window	F	p	F	p
Panel A: IMDb				
Days −30 to −9	1.386	0.109	0.907	0.586
Days −14 to −9	2.001	0.062	1.290	0.258
Panel B: Letterboxd				
Days −30 to −9	3.527	<0.001	0.738	0.803
Days −14 to −9	1.401	0.211	0.585	0.742

Notes: This table reports joint F-tests of the null that the pre-onset event-study coefficients in the window of each row equal zero, from the stacked DiD specification, on IMDb (Panel A) and Letterboxd (Panel B). The omitted period is day −8, the last day before the treatment onset. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes (reviews); the implied percent effect is $100(e^{\beta} - 1)$.

Appendix A Additional Results

In this appendix, I report the additional results and exhibits referenced in Sections 4 and 5. In Appendix A.1, I report the pre-trend tests and the onset-sensitivity estimates for both platforms. In Appendix A.2, I report the between-film content-gradient estimates, including the coefficients behind Figure 4 and the recommendation-valence slopes on volume. In Appendices A.3 to A.5, I provide the robustness checks described in Section 5.

A.1 Pre-Trends and Onset Sensitivity

This section reports the tables behind the pre-trend discussion in Section 4. In Table A.1, I report the joint F-tests on the pre-onset event-study coefficients for both platforms. Table A.2 reports ATTs and pre-trend tests under onsets of 0 (treatment at the air date), −7 (primary), and −14, with reference periods −1, −8, and −15 respectively. On IMDb, dating treatment from the air date (onset 0) leads to a rejection of parallel pre-trends for the volume outcome because it includes the anticipatory week in the baseline. Under the −14 onset, the volume ATT is 12.9% compared to 16.4% under the primary onset, the rating ATT is nearly unchanged, and the volume pre-trend test does not reject. On Letterboxd, the volume pre-trend test rejects under every onset (see the discussion in Section 4.2), while the rating test rejects only when treatment is dated from the air date. The −7 and −14 onsets produce similar Letterboxd ATTs for both outcomes.

Table A.2: Sensitivity to Treatment Date

	Onset = 0 (ref. −1)	**Onset = −7** (primary; ref. −8)	Onset = −14 (ref. −15)
Panel A: IMDb			
Average treatment effect			
Rating volume (daily new votes)	0.1271***	0.1517***	0.1212***
	(0.0290)	(0.0285)	(0.0275)
Average rating	-0.0026***	-0.0021***	-0.0019***
	(0.0007)	(0.0007)	(0.0006)
Pre-trend joint test (p-value)			
Rating volume (daily new votes)	<0.001	0.109	0.156
Average rating	0.580	0.586	0.653
Observations			
Rating volume (daily new votes)	111,568	111,568	111,568
Average rating	117,522	117,522	117,522
Treated films			
Rating volume (daily new votes)	540	540	540
Average rating	560	560	560
Panel B: Letterboxd			
Average treatment effect			
Rating volume (daily new reviews)	0.1640***	0.8237***	0.8491***
	(0.0387)	(0.0588)	(0.0603)
Average rating	-0.0509***	-0.1127***	-0.1079***
	(0.0183)	(0.0210)	(0.0236)
Pre-trend joint test (p-value)			
Rating volume (daily new reviews)	<0.001	<0.001	0.001
Average rating	0.006	0.803	0.677
Observations			
Rating volume (daily new reviews)	63,560	63,560	63,560
Average rating	31,929	31,929	31,929
Treated films			
Rating volume (daily new reviews)	261	261	261
Average rating	261	261	261

Notes: This table reports stacked DiD average treatment effects on IMDb (Panel A) and Letterboxd (Panel B) under three treatment-onset assumptions, with event-study reference periods −1, −8, and −15 respectively. Within each panel, the pre-trend rows report the p-value of a joint F-test on the pre-onset event-study coefficients. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes (reviews); the implied percent effect is $100(e^{\beta} - 1)$. Standard errors are clustered at the film level and reported in parentheses.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.2 Between-Film Content Gradient

In Section 4.4, I present the quality-assessment slopes for ratings in Figure 4. I report the between-film content gradients from Equation (6) in Table A.3, pairing each sentiment measure with the margin the framework assigns it (Section 3.5): recommendation valence on rating volume in columns (1) and (2), and quality assessment on the average rating in columns (3) and (4), each on IMDb and then Letterboxd. In every column of Table A.3, the Coverage row is the treatment effect at a score of zero, and the Coverage × score row is the between-film slope, β_2. The table adds to Figure 4 the coefficients behind its two panels and the two recommendation-valence slopes on volume. The rating slopes are positive on both platforms. Figure A.1 redraws Figure 4 with the rating effect also estimated separately within each tercile of the quality assessment, taken over the covered films in each platform's content-gradient sample.

Prediction 1 (Section 2) states that coverage should raise volume regardless of the hosts' opinions of the film. That is ultimately a statement about the sign of the treatment effect across the range of observed scores, not just about its slope, and so a nonzero slope would threaten the prediction only if the effect turned negative somewhere in that range. Table A.3 shows that the Letterboxd recommendation-valence slope for the volume outcome is negative, but the implied effect at the top of the score range remains positive, so coverage raises volume at every observed score on both platforms.

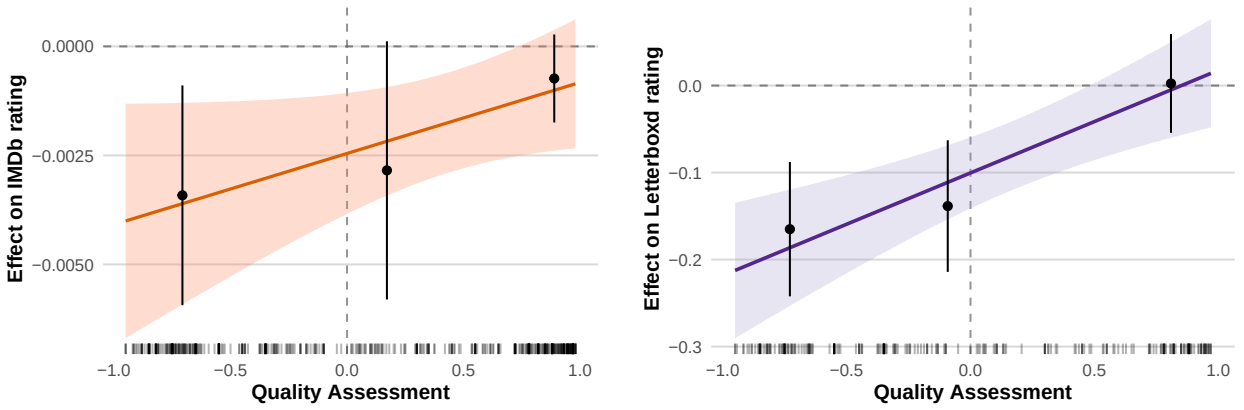
A.3 Shifted-Date Placebo

In Table A.4, I report results of a placebo test where I shift the treatment date for each film earlier by 60 days. Each cohort keeps its covered film and its matched controls, and its air date moves earlier by 60 days, so that the event window runs from 90 to 30 days before the actual air date. I call this shifted treatment date the film's "placebo date." I then estimate Equation (5) as in the main analysis around the placebo dates, with the event window, the reference period, corrective weights approach, and the clustering unchanged. The matched controls are the same as in the primary analysis. The "Actual air dates" column reports the main estimates of Section 4 for comparison. Neither placebo effect is distinguishable from zero. The volume placebo is 0.5% versus 16.4% at the actual air dates, and the rating placebo is 0.0000.

Table A.3: Between-Film Content Gradient

	(1)	(2)	(3)	(4)
	Rating volume		Average rating	
Score	IMDb Rec. valence	Letterboxd Rec. valence	IMDb Quality	Letterboxd Quality
Coverage	0.1594***	0.7618***	-0.0025***	-0.1004***
	(0.0503)	(0.0659)	(0.0007)	(0.0208)
Coverage × score	-0.0026	-0.1922*	0.0016*	0.1180***
	(0.0632)	(0.1090)	(0.0009)	(0.0305)
Observations	107,830	59,651	115,054	29,847
Covered films with a score	517	255	546	258

Notes: This table reports stacked DiD estimates of Equation (6), each column a separate regression at the film-day level. Columns (1) and (2) interact coverage with recommendation valence and take rating volume as the outcome; columns (3) and (4) interact coverage with quality assessment and take the average rating as the outcome. Columns (1) and (3) use the IMDb sample and columns (2) and (4) the Letterboxd sample. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes (reviews); the implied percent effect is $100(e^{\beta}-1)$. Standard errors are clustered at the film level and reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.



(a) IMDb

(b) Letterboxd

Figure A.1: Treatment Effect by Quality Assessment, Linear Fit and Terciles

Notes: This figure adds to Figure 4 stacked DiD estimates of the effect on the average rating within terciles of the quality assessment, plotted at each tercile's mean score with 95% confidence intervals. The tercile estimates come from one regression that replaces Coverage and Coverage × quality in Equation (6) with Coverage interacted with an indicator for each tercile, on the same sample. A joint Wald test of linearity, which adds the second- and third-tercile interactions to Equation (6) and tests that both are zero, has a p-value of 0.90 on IMDb and 0.53 on Letterboxd. Standard errors are clustered at the film level.

Table A.4: Placebo Test: Shifted Air Dates (IMDb)

	Actual air dates	Shifted −60 days
Rating volume (daily new votes)	0.1517***	0.0053
	(0.0285)	(0.0240)
Average rating	-0.0021***	0.0000
	(0.0007)	(0.0005)
Treated films (volume)	540	543
Treated films (rating)	560	561

Notes: This table reports stacked DiD average treatment effects on IMDb. The "Actual air dates" column reports the main estimates, and the "Shifted −60 days" column re-estimates the same specification on the same covered films and matched controls with every air date moved earlier by 60 days. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes; the implied percent effect is $100(e^{\beta} - 1)$. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.4 Data Coverage

Table A.5 reports the main IMDb specifications re-estimated on the treatment events that air on or after January 21, 2023, the date from which the ratings record is daily, next to the full-sample estimates (Section 5.2). The restriction does not change the results. Coverage raises the volume of new votes by 19.3% in the daily-coverage sample, against 16.4% in the full sample, and the effect on the average rating is -0.0020 against -0.0021. The volume pre-trend test does not reject at the 5% level in either sample, though in the daily-coverage sample its p-value falls just below the 10% threshold ($p = 0.099$, versus $p = 0.109$ in the full sample).

A.5 Functional Form of Rating Volume

As a comparison to the primary PPML estimate of the coverage effect on daily new votes (reviews), I estimate a linear model with a $\log(1 + x)$ transformation as the dependent variable. I present the results in Table A.6 with IMDb in Panel A and Letterboxd in Panel B. Both use the same fixed effects and clustering. The two estimators agree in sign and significance on both platforms, and on IMDb in magnitude as well. On Letterboxd, the log OLS coefficient is a little over half the PPML coefficient, a gap I discuss in Section 5.

Table A.5: Restricting to the Period with Daily Data Coverage (IMDb)

	(1) Full sample	(2) Daily coverage only
Coverage effect on rating volume (daily new votes, PPML)	0.1517***	0.1766***
	(0.0285)	(0.0325)
Coverage effect on average rating	-0.0021***	-0.0020***
	(0.0007)	(0.0007)
Observations (volume)	111,568	106,583
Observations (rating)	117,522	111,910
Treatment events (volume)	548	319
Treatment events (rating)	568	319
Pre-trend p (volume)	0.109	0.099
Pre-trend p (rating)	0.586	0.983

Notes: This table reports stacked DiD average treatment effects on IMDb for the full sample and for the subsample of treatment events airing on or after January 21, 2023, the date from which the Historical IMDb Ratings service records each title page daily. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes; the implied percent effect is $100(e^{\beta} - 1)$. Pre-trend p-values are from a joint F-test on the pre-onset event-study coefficients. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix B TWFE Results

This appendix reports two-way fixed effects (TWFE) counterparts to the film-level specifications on both platforms. As discussed in Section 4.1, the TWFE estimator's implicit comparisons can bias its estimates under staggered treatment timing (Goodman-Bacon, 2021), which is why the stacked estimator is my primary choice. In order to isolate the differences across estimators, the TWFE specification estimated here uses the same films and treatment events as in the stacked DiD approach from Section 4.1. In practice, the two estimators agree in sign and significance in every average treatment effect this appendix reports, and in magnitude for each of them except the Letterboxd volume effect, which is larger under TWFE (Table B.1). Other than that estimate, the TWFE bias appears to be small in this setting.

B.1 Main Results

Table B.1 reports the TWFE average treatment effects, with IMDb in Panel A and Letterboxd in Panel B. On both platforms, I pool the event windows, place each film-day on its calendar date, and estimate the specification with film and date fixed effects. A control film matched to more than

Table A.6: Functional Form of Rating Volume

	(1) Poisson PPML	(2) OLS
Panel A: IMDb		
Outcome	Daily new votes	Log(1 + daily new votes)
Coverage effect	0.1517***	0.1768***
	(0.0285)	(0.0168)
Implied percent effect	16.4%	–
Observations	111,568	115,781
Treated films	540	559
Panel B: Letterboxd		
Outcome	Daily new reviews	Log(1 + daily new reviews)
Coverage effect	0.8237***	0.4517***
	(0.0588)	(0.0251)
Implied percent effect	127.9%	–
Observations	63,560	65,880
Treated films	261	261

Notes: This table reports the stacked DiD coverage effect on rating volume under two functional forms, on IMDb (Panel A) and Letterboxd (Panel B). Column (1) estimates the volume effect by Poisson pseudo-maximum likelihood on the daily count of new votes (reviews), and column (2) by OLS with $\log(1 + x)$ of the same count as the outcome; the implied percent effect is $100(e^{\beta} - 1)$. Both columns use the same fixed effects and clustering, and each column's corrective weights are computed on its own sample. The implied percent effect is reported for column (1) only: when the outcome is often zero, the log OLS coefficient is not a percent effect (Chen and Roth, 2024). The Poisson sample is smaller because the estimator drops fixed-effect cells in which the outcome is always zero. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.1: TWFE Main Results: Effect of Podcast Coverage on Rating Activity

	(1) Rating volume	(2) Average rating
Panel A: IMDb		
Coverage	0.1857***	-0.0017**
	(0.0400)	(0.0007)
Observations	116,441	120,999
Treated films	540	560
Panel B: Letterboxd		
Coverage	1.0477***	-0.1117***
	(0.0715)	(0.0187)
Observations	65,577	39,136
Treated films	261	261

Notes: This table reports average treatment effects from the TWFE specification, with film and calendar-date fixed effects. Both panels are at the film-day level, pooled across event windows on calendar dates. Panel A uses the IMDb sample: rating volume is daily new votes, and the average rating is the displayed average on a 1–10 scale. Panel B uses the Letterboxd sample: rating volume is daily new reviews, and the average rating is the daily mean of new reviews' star ratings on a 0.5–5 scale. Volume effects are estimated by Poisson pseudo-maximum likelihood; the implied percent effect is $100(e^{\beta} - 1)$. On IMDb, the volume column uses the films and treatment events of the stacked volume sample, and the rating column those of the stacked rating sample. Standard errors are clustered at the film level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

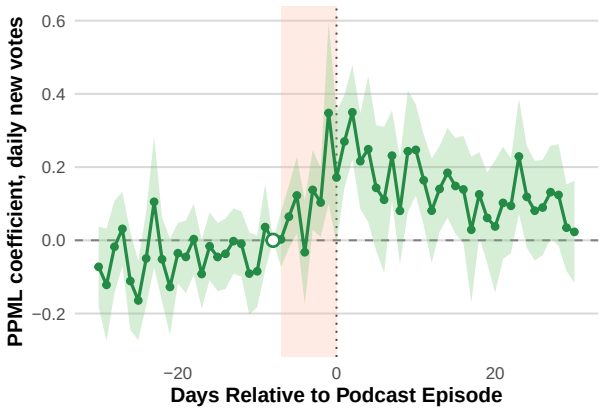
one treated film enters once per calendar day. On Letterboxd, coverage raises daily new reviews by 185.1% under TWFE against 127.9% under stacking, and the effect on the daily mean rating is -0.112 stars against -0.113.

B.2 Event Studies

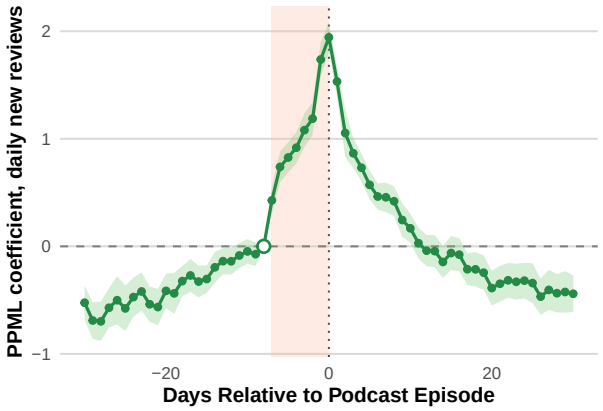
I present TWFE event-study figures in Figure B.1. They are consistent with the stacked DiD results presented in Section 4. There is a sharp spike in new ratings that begins about a week before air, and a decline in the average rating following the release of the episode. The Letterboxd volume pre-trend test rejects (p <0.001), as it does in the stacked specification, reflecting the anticipatory drift discussed in Section 4, while the Letterboxd rating pre-trend test does not ($p = 0.318$).

B.3 Sentiment

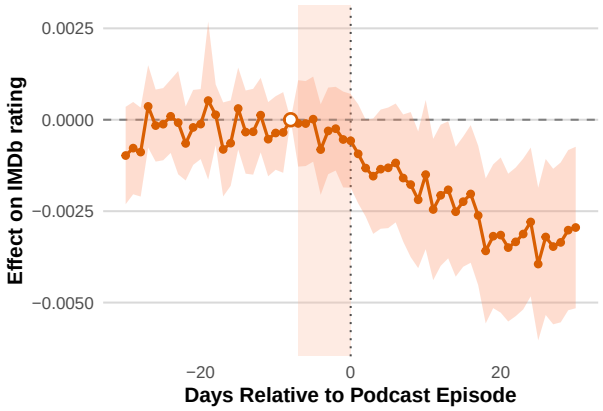
Figure B.2 plots the TWFE sentiment interactions on IMDb. As in the stacked results, the quality-assessment slope on the rating is positive and the recommendation-valence slope on volume is flat.



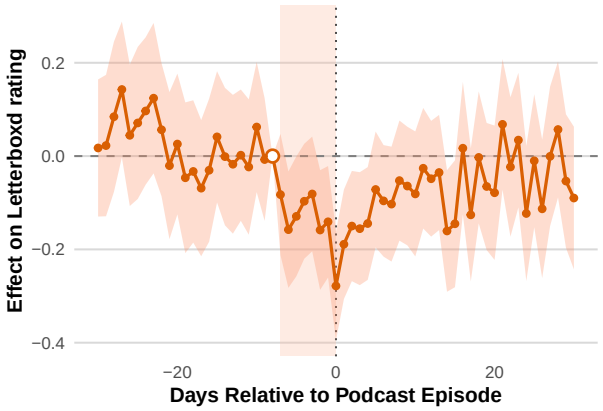
(a) IMDb: Daily New Votes



(b) Letterboxd: Daily New Reviews



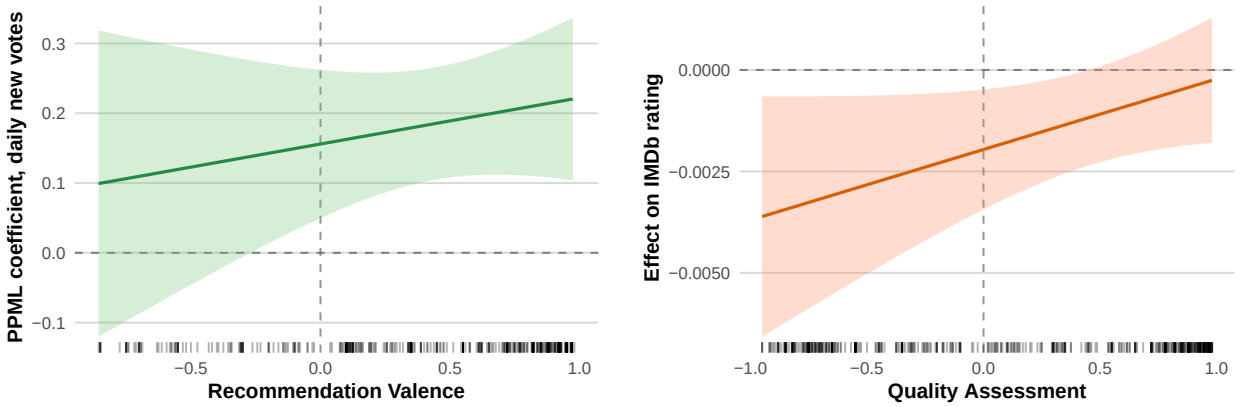
(c) IMDb: Average Rating



(d) Letterboxd: Daily Mean of New Ratings

Figure B.1: TWFE Event Studies: Rating Volume and Average Rating

Notes: This figure plots point estimates and 95% confidence intervals from the TWFE specification, with film and calendar-date fixed effects. Panels (a) and (b) show rating volume: daily new votes on IMDb and daily new reviews on Letterboxd. Volume effects are estimated by Poisson pseudo-maximum likelihood; the implied percent effect is $100(e^{\beta} - 1)$. Panels (c) and (d) show the average rating: the displayed average on IMDb and the daily mean of new reviews' star ratings on Letterboxd. The omitted period is day -8. The shaded region from the treatment onset to the air date is the anticipatory window, and the dotted vertical line marks the air date.



(a) Recommendation Valence × Rating Volume

(b) Quality Assessment × Average Rating

Figure B.2: TWFE: Treatment Effect by Sentiment (IMDb)

Notes: This figure plots the linear interaction estimates of τ(recommendation valence) on daily new votes (panel (a)) and of τ(quality assessment) on the IMDb rating (panel (b)), using the TWFE specification. Shaded areas are 95% confidence bands, and the marks along the horizontal axis show the covered films' scores. Volume effects are estimated by Poisson pseudo-maximum likelihood on the daily count of new votes; the implied percent effect is $100(e^{\beta} - 1)$.

Appendix C Transcript Scoring Prompt

This appendix documents the procedure used to score episode transcripts (Section 3.5). Each episode–film pair is scored with Anthropic's `claude-sonnet-4-6` model through the Batch API, with structured (tool-call) output so the model returns a fixed JSON schema. The prompts reproduced below were drafted with the assistance of Claude and revised by hand. Each transcript goes through two passes.

The *scoring pass* determines the podcast's recommendation valence and quality assessment scores. The *evidence pass* identifies supporting quotes and constructs the discussion share, an enthusiasm score, and a set of flags.[27] The scoring pass is run eight times at a sampling temperature of 0.7. I define each film's recommendation valence and quality assessment as the average of the eight draws. In this pass, I do not use step-by-step reasoning, since that would collapse the eight draws toward a single value. Each draw is given an identical rubric that describes the ends of each scale in words instead of fixed numeric anchors, and instructs the model to judge the discussion of the film itself, not simply the tone of the conversation. This guards against, for example, cases

[27] The flags record whether the episode discusses multiple films substantially, whether the film is barely mentioned, whether the transcript has quality problems, and—separately for recommendation valence and quality assessment—whether the hosts make no evaluative statement on that dimension.

where the hosts might excitedly discuss their dislike of a film.

The evidence pass is run once at temperature zero. I use the discussion share when constructing the sample. When it falls below 0.1, or when the film is flagged as barely mentioned, a film's scores are treated as missing, as are the scores for any dimension flagged as lacking an evaluative statement. I use the multiple-film and transcript-quality flags for diagnostic purposes, and they do not play a role in the construction of my analysis sample.

Every request shares the same system prompt, followed by the film's title and release year and the full episode transcript. The user message provides the task instruction for each pass.

System Prompt

```
You are a research assistant analyzing how a podcast episode discusses a
specific movie. Below are the movie's identity and the full episode transcript.
A separate instruction message will tell you exactly what to produce; apply the
principle below in whatever you are asked to do.

## Core principle: judge the FILM, not the CONVERSATION

Score the hosts' assessment OF THE MOVIE -- not the mood of the conversation.
These are entertainment shows; hosts are often funny, energetic, and warm even
about films they think are bad. Judge what they say about the movie's merits
and whether they would steer a listener toward it, ignoring how lively or
pleasant the discussion itself is.

Examples:
  - Hosts laugh hard and clearly enjoy recording, but call the film incompetent
    and say it only "works" as something to mock -> LOW quality_assessment,
    regardless of the upbeat mood.
  - Hosts are calm and dry but describe the film as a masterwork they would
    urge anyone to see -> HIGH recommendation_valence and quality_assessment.
  - Warm banter, inside jokes, and host chemistry -> IRRELEVANT to both scores.

## Do NOT use as evidence (for either dimension)
```

Laughter, banter, jokes, the hosts' energy or enthusiasm, host chemistry, friendly rapport, or how much fun the hosts are having. Those describe the conversation, not the film.

Scoring Pass (Run Eight Times)

Produce exactly two scores for the movie above, each a real number from -1 to +1.

recommendation_valence -- how strongly the hosts steer listeners toward or away from watching the film.
  near -1 : actively warn listeners off ("don't bother", "skip it").
  around 0: no clear steer, or genuinely mixed.
  near +1 : urge listeners to watch ("you have to see this").

quality_assessment -- how good or bad the hosts judge the film to be as a piece of filmmaking, independent of whether they recommend watching it.
  near -1 : describe it as badly made on most dimensions.
  around 0: mixed or middling.
  near +1 : describe it as excellent, among the best.

These can diverge: hosts may push listeners toward a film they consider terrible because it is entertainingly bad.

Give your best point estimate using whatever value the evidence supports -- ANY number in [-1, 1], not only round values like 0.5 or 0.8.

Return the two scores using the score_film tool. Do NOT explain, justify, or add any commentary -- return only the tool call.

Evidence Pass (Run Once)

Gather textual evidence and record coverage and data-quality flags for the movie above. You are NOT producing the numeric sentiment scores here -- a separate process does that.

Procedure:

1. Extract 2-5 short verbatim quotes in which the hosts evaluate whether the FILM is worth watching (recommendation), and 2-5 short verbatim quotes in which they evaluate the FILM's quality or craft. Every quote must be about the movie itself -- not the hosts, the podcast, or unrelated tangents. If the hosts make no evaluative statement on a dimension, return an empty list and set the matching insufficient-evaluation flag.
2. Estimate discussion_share: the fraction of the episode actually spent on this movie (exclude ads, intros, host banter, and discussion of other movies).
3. Estimate enthusiasm: the overall energy and engagement of the discussion. (This one IS about the conversation's energy, not the film's quality.)
4. Set the flags honestly.

Return everything via the extract_evidence tool.

Transcript (Appended to the System Prompt)

Movie
Title: {title} ({year})
IMDb ID: {tconst}
Podcast: {podcast_name}

--- TRANSCRIPT ---

{full_text}