The Good, The Bad and The Picky:

Consumer Heterogeneity and the Reversal of Product Ratings

Tommaso Bondi, Michelangelo Rossi and Ryan Stevens*

March 28, 2024

Abstract

We study the impact of consumer heterogeneity on online ratings. Consumers differ in their experience, which can affect both their choices and ratings. Thus, biases in average ratings can arise when the opinions of experienced and novice users are aggregated. We first build a two-period model to characterize the biases' drivers and consequences. We test our theory combining data from IMDb and MovieLens, two well-known movie ratings platforms. We proxy users' experience with the total number of ratings posted on the platforms. First, using external measures of quality, such as the Academy awards and nominations, we show that, on both platforms, experienced users, on average, rate movies of higher quality compared to novices. Moreover, they post more stringent ratings than novices for over 98% of movies. Combined, these imply a compression in aggregate ratings, and thus a bias against high quality movies. We then propose a simple, fixed-point algorithm to de-bias ratings. Our de-biased ratings demonstrate the presence of ranking reversals for over 8% of comparisons in our sample. As a result, our de-biased ratings better correlate with external measures of quality.

*T. Bondi: Cornell University, and CESifo. E-mail: tbondi@cornell.edu. M.Rossi: Télécom Paris, Institut Polytechnique de Paris, and CESifo. E-mail: michelangelo.rossi@telecom-paris.fr. R. Stevens: Ramp Corp. E-mail: ryan.louis.stevens@gmail.com.

1 Introduction

Online consumer ratings have become a ubiquitous driver of choice. But, to what extent can we trust their informational content? On one hand, on all well known online ratings platforms, the ratings of extremely experienced users, who sometimes post thousands of ratings over many years, are aggregated with those of a long tail of inexperienced users who only leave a handful of ratings. On the other hand, ratings can reflect characteristics of their writers just as much as of what is being reviewed. This can introduce biases in aggregate ratings whenever individuals' characteristics correlate with their choices.

Typically, such self-selection biases stem from differences in *taste*: when products are horizontally differentiated, reviews reflect product-consumer fit just as much as product quality *per se*. In this paper, we are interested in a complementary and – we argue – equally important form of self-selection, arising even when all consumers would rank all products equally.

The mechanism we propose is simple: consumers differ in their experience, and experience affects choice: experience users, for instance, may have higher expertise, which leads them to observe better quality signals and to choose, on average, higher quality products. Moreover, experienced users might also rate products differently from novices. For example, their ratings may be more diagnostic of quality or simply more (or less) stringent than those of novices for each quality level.

When choice and rating heterogeneity are correlated, aggregating ratings from different user types can therefore result in an *"apples-with-oranges"* comparison between the average product ratings of different products. Moreover, unlike taste-based self-selection, the mechanism we propose biases ratings comparisons even for products that are horizontally very similar.

To illustrate this phenomenon, consider the following example: Experienced and novice users choose between two movies, A and B. On average, experienced users rate A as 7 and B as 6, while novices rate them 8.5 and 8, respectively. Moreover, 90% of experienced users choose movie A: their expertise is instrumental in identifying the superior option. Novices, on the other hand, pick A and B randomly.Once reviews are aggregated, movie A and B's scores are given by the weighted average of ratings from the two groups, that is, $\mathscr{R}(A) = \frac{0.9 \cdot 7 + 0.5 \cdot 8.5}{0.9+0.5} = 7.53$ and $\mathscr{R}(B) = \frac{0.1 \cdot 6 + 0.5 \cdot 8}{0.1+0.5} = 7.66$. That is, while everyone agrees that $Q(A) > Q(B)$, $\mathscr{R}(B) > \mathscr{R}(A)$ due to the less stringent reviews product B receives on average.[1] Future generations of users who employ these ratings in their decisions will be worse off than if ratings did not exist in the first place.

[1]This examples makes apparent a conceptual link between our model and the Simpson's Paradox (Blyth, 1972). We thank Nikhil Garg for suggesting this interpretation.

We first build a theoretical model to characterize the drivers and consequences of this phenomenon. In our model, there are two generations of users. In both periods, different user types observe signals of (potentially) heterogeneous precision about the quality of two products. Moreover, first generation users leave a rating to the product they choose. Our model remains agnostic on the possibly heterogeneous ways in which experienced and novice users map quality into ratings. Each rating also contains an idiosyncratic shock which, combined with the number of ratings, determines the ratings' aggregate precision, and thus their impact on second generation users' posteriors and choices.

We start by highlighting our model's main results. First, if users' ratings respond equally to quality increases, a rating compression - i.e., rating differences that understate underlying quality differences - occurs if and only if experienced users' ratings are more stringent. When experienced users are much more stringent than novices, rating differences not only compress quality differences, but reverse them altogether, a phenomenon we call ranking reversal. In addition, whenever novice users are less sensitive to quality than experienced users, rating compressions and ranking reversals become more prevalent.

Second, we study the consequences of both compressions and reversals on second period users' choices. We assume second period users learn naïvely from relative ratings, that is, they take them at face value. This is well justified in our context, since correcting the ratings requires knowledge of each product's quality – which is the unknown object of learning in the first place.[2] We show that compressed ratings can lower consumer welfare, and that they are most likely to do so when their precision is intermediate: when ratings are very noisy, users ignore them; and when they are very precise, they are useful since they correctly rank the two products (albeit understating their quality difference). With ranking reversals, however, the conclusions become starker: reversals hurt all consumers, the more so the more precise the ratings.

We then discuss platform design remedies. We start by showing that a commonly employed practice, that of overweighting the ratings of the more experienced users, can backfire. To see why, notice that whenever they are more stringent, experienced users impose a penalty on high quality products. Overweighting their ratings increases this penalty. However, less intuitively, we show that this is not the case whenever the proportion of ratings from experienced users is high to begin with: in this case, overweighting them makes the crowd of users more homogeneous, leading the platform's ratings to be approximated by those of the experienced users ones only, and thus mitigating the "apples-with-oranges" problem.

However, the platform can do better. The key observation to this end is that however pervasive, this type

[2]It also requires knowledge of the proportion of experienced and novice users, as well as their choice and ratings patterns.

of self-selection bias allows for a straightforward correction. We exploit the history of individual ratings and compute a stringency score for each user. Then, we subtract user-specific stringency from each of the user's ratings. We then use the corrected individual ratings to updated average product ratings. We iterate this process until it converges, that is, until individual stringencies and product ratings are self-confirming. This allows the platform to not discard any rating, but rather make a better use of the existing ones by simply "leveling up the playing field", that is, by holding each product to the same standard.

We empirically substantiate the presence and severity of these biases – as well as its drivers – by studying consumer movie ratings. In particular, we scraped detailed aggregate data for over 9,000 movies from IMDb, the most popular movie rating platform in the world,[3] and complemented it with a massive (over 25 million reviews and 32,000 individual users) individual-level rating dataset from MovieLens, a platform for movie discovery and personalized recommendations created by GroupLens, a research group within the University of Minnesota's Computer Science department.

The combination of these two datasets offers unique advantages for the purposes of our study. First, the industry itself is convenient: movies are uniformly priced, which allows us to focus on the relationship between quality and ratings without having to consider how price shapes both choices and ratings.

Second, and more importantly, the two platforms offer complementary data. IMDb divides its users between Top1000 (those who have posted the most ratings on the platform; an extremely elite group, given IMDb's over 200 million registered users) and Non-Top1000, and, for each movie, displays both the number and average of ratings from both user groups. This is the perfect setting for us to study both choice and rating heterogeneity. However, IMDb does not allow us to see individual users' history, making it impossible to apply our de-biasing algorithm. Thankfully, MovieLens is the perfect setting for this, as it allows us to track over $32,000$ users over time. Moreover, it allows us to double-check all of our IMDb findings, and importantly, to do so with a less dichotomic and more time-varying definition of individual experience.

The two platforms tell us a similar story in terms of both choice and rating heterogeneity. Experienced users watch, on average, higher-quality movies. To proxy movie quality, we employ external sources such as Academy Awards nominations and awards, as well as a host of the most relevant festival and industry awards around the world and critics' reviews aggregated by Metacritic. Moreover, experienced users are significantly more stringent than novice users in their ratings. This is not only true on average, but the result

[3] As of 2022, IMDb has over 83 million registered users and 200 million unique monthly visitors; as of January 2024, it is one of the top 65 most visited websites worldwide.

holds for 98% of movies in our sample. This rating gap is statistically and economically significant: for a given movie on IMDb, experienced users' ratings are, on average, over half-star lower than novices'.

Two facts are worth pointing out to rule out taste differences between experienced and novice users as the main explanation for our observed patterns. First, experienced and novice users like the same movies, as shown by the fact that their ratings are very highly correlated (0.89). Second, the difference in ratings between experienced and novice users is stable across genres: thus, it is not the case that experienced users simply like certain niche genres more than novice, and dislike mainstream ones such as Action or Thriller. Rather, and more simply, experienced users uniformly rate nearly every movie lower than novices do.

The combination of experienced users' quality-based self-selection and their stringent rating behavior implies that aggregate ratings are compressed, and therefore penalize high quality movies compared to their inferior alternatives.

Last, we apply the aforementioned algorithm to MovieLens ratings. Without any assumptions on the relation between experience and choice or rating behavior, we let our algorithm recursively compute each movie corrected rating as well as each individual stringency. This correction is appealing in that it does not require us to take a stance on which ratings (or which users) are more or less accurate, nor on which movies are of higher or lower quality. We then use the de-biased ratings and user stringencies to offer an assumption-free confirmation of our previous findings: more experienced users, on average, are more stringent and choose movies with higher de-biased ratings (a proxy for quality). As a natural consequence, our algorithm shows that movies with higher de-biased ratings face, on average, a more stringent crowd, and are thus penalized on MovieLens. Crucially, our de-biased aggregate ratings display ranking reversals for around 8% of the movie pairs in our dataset. Last, they better correlate with external measures of quality, such as Academy nominations and awards and reviews by critics.

The rest of the paper is structured as follows: Section 1.1 surveys the literature; Section 2 presents our theoretical model; Section 3 briefly draws parallels between our theory and empirics; Section 4 and 5 describe the data and the empirical analysis; Section 6 presents our de-biasing algorithm and its impact on aggregate ratings. Section 7 concludes.

1.1 Related Literature

This paper adds to a large and highly multidisciplinary body of research focusing on online ratings and reviews.[4] A first strand of this literature has focused on quantifying the impact of ratings on both choice and welfare. Seminal work by Chevalier and Mayzlin (2006), and more recently by Anderson and Magruder (2012), Yoganarasimhan (2013), Luca (2016), Lewis and Zervas (2016), Reimers and Waldfogel (2021), Farronato and Zervas (2022), Fang (2022) find sizable causal impact.

A second strand of research documents, theoretically and empirically, the nature of ratings, as well as their systematic biases. Biases in ratings can result from both sellers' strategic behavior (Chevalier, Dover and Mayzlin (2014), Anderson and Simester (2014), Luca and Zervas (2016), He, Hollenbeck and Proserpio (2022), Akesson, Hahn, Metcalfe and Monti-Nussbaum (2023)), and consumer taste-based self-selection (Li and Hitt (2008), Godes and Silva (2012), Besbes and Scarsini (2018), Acemoglu, Makhdoumi, Malekian and Ozdaglar (2022), Bondi (2023)).

Brandes, Godes and Mayzlin (2022) focus on the fact that selection happens at the ratings phase, conditional on choice: only those with extreme opinions about a product review it. However, Schoenmueller, Netzer and Stahl (2020) show that extremity bias does not uniformly affect all platforms, and that "information platforms" with a larger scale for ratings, like MovieLens and IMDb, exhibit less polarity imbalance.[5]

Related to our attempts to reconcile consumer-generated ratings with other, non-consumer-generated proxies of quality, De Langhe, Fernbach and Lichtenstein (2016) document a low correlation between consumer and professional reviewers' opinions. Importantly, this holds even in markets without substantial product differentiation.[6] Moreover, consistent with our modeling assumptions, De Langhe et al. (2016) point out that consumers do not appear to internalize biases in online reviews, and instead take them at face value.[7] We therefore carry the analysis under the assumption that the average ratings posted by the platform are the sole drivers of social learning.[8] This is in line with Acemoglu et al. (2022), who stress how *"It is important to move beyond Bayesian learning and investigate what types of rating systems robustly aggregate*

[4] For overviews of this topic, see Cabral (2012) and Tadelis (2016).

[5] Our data confirm their findings – for a detailed discussion, see Section 4.4.

[6] Winer and Fader (2016) argue that low correlation is neither surprising nor necessarily problematic: less correlated sources of information are jointly more informative. This is certainly true when low correlation is due to, for instance, taste differences. Our paper, on the other hand, suggests it might be due to problematic aggregation of consumer ratings.

[7] Ours is, of course, not the first model of naïve social learning. In fact, a considerable portion of the existing social learning literature deals with boundedly rational agents (for instance, in Ellison and Fudenberg (1995), agents simply count previous adopters). This is motivated by both realism and analytical tractability. Moreover, Esponda (2008) and Spiegler (2016) focus on the misunderstanding of selection effects, albeit not directly applied to social learning contexts.

[8] We note that, implicitly, this is also the approach of a large majority of the empirical literature on the topic.

information when agents use simple learning rules" (such as comparing averages in ratings).[9]

By focusing on the relationship between experience on the platform and choice, we implicitly highlight the interplay between experience and expertise. How are they related? A large literature in economics defines experts as those with more accurate information (Crawford and Sobel (1982); Krishna and Morgan (2001); Gerardi and Yariv (2008)). The seminal work of Arrow (1962) poses that we *"learn by doing"*, thus implying a positive correlation between the two measures.[10] Alba and Hutchinson (1987) highlight that the two constructs are distinct, but closely related.[11] In this paper, we mostly refer to experienced users and novices, rather than experts and non-experts, because we can easily measure experience (the number of ratings left by each user), but not expertise. Nevertheless, we exploit the correlation between these two constructs, and our model assumes – in line with the economic definition of expertise – that experienced users observe more accurate signals of quality, leading them to consume higher quality on average (an assumption we test and confirm empirically).

Turning to the relationship between experience/expertise and rating behavior, LaTour and Deighton (2019) stress the lack of research discussing how, and whether, expertise shapes consumers' hedonic gratification. Similarly, Rocklage, Rucker and Nordgren (2021) point out that Alba and Hutchinson (1987) never mention the word "hedonic", and solely focus on the relationship between expertise and choice.

LaTour and Deighton (2019) argue that *"evidence suggests that consumers seek to become more expert about hedonic products to enhance their enjoyment of future consumption occasions"*. This is at odds with Rocklage et al. (2021), who find that expertise leads to "emotional numbness". While our goal is not to unpack the psychological mechanisms underpinning differences in ratings between experienced and novice users, we do find that experienced users leave systematically lower ratings, and that users become harsher as they accumulate ratings on the platform.

Nguyen, Wang, Li and Cotte (2021) experimentally show that expertise moderates extremity bias, and argue that this is due to experts considering more dimensions in their ratings. As a result, they argue, experts leave lower (higher) ratings for high (low) quality products. Instead, we find that only 1.5% of users in our samples tend to leave u- or j-shaped ratings, and that the proportion of experienced and novice users who do so is statistically indistinguishable. Moreover, experienced users are more stringent than novices for each

[9] For an overview of the importance of accounting for behavioral biases in strategic models, see Narasimhan, He, Anderson, Brenner, Desai, Kuksov, Messinger, Moorthy, Nunes, Rottenstreich, Staelin, Wu and Zhang (2005), as well as Spiegler (2011).

[10] Similarly, Ericsson, Krampe and Tesch-Römer (1993) stress the importance of deliberate practice to build expertise.

[11] See also Johnson and Mervis (1997) on the positive correlation between the two.

quality level; if anything, we find a small decline in this difference for high quality products, consistent with experienced users' ratings being more diagnostic of quality.

Importantly, Nguyen et al. (2021) also discuss the relation between experience and expertise in the context of online reviews, stressing that platforms such as Yelp and Amazon effectively equate the two in their definition of their top reviewers. They argue that their findings provide evidence for a positive correlation between the two measures: *"[...] online reviewing experts, as designated by many online review platforms, largely exhibit features of expertise, including a greater degree of elaboration, and greater category knowledge."* While we focus on numerical ratings rather than textual reviews, we also find that experienced users are also more expert, as they consume (or at least rate) higher quality products on average.

Our findings are also closely related to the literature on scale usage bias (Greenleaf (1992), Rossi, Gilula and Allenby (2001)). Using consumer satisfaction survey data, Rossi et al. (2001) propose a Bayesian hierarchical approach to correct for the bias, and show that their corrected measure better correlates with purchase intentions. Our correction is similar in spirit, but methodologically different. While they assume priors for both the scale and the bias parameters, we initialize the process at their empirical value using the original platform ratings, and then leverage the movies-users network to iterate forward.

2 Model

2.1 Overview

We begin by theoretically modeling the properties and consequences of ratings with consumer heterogeneity. Following a rich literature in marketing, operations, and economics (see, for example, Sun, 2012; Papanastasiou and Savva, 2017; Fainmesser, Olié Lauga and Ofek, 2021), we study the causes and effects of the aggregation bias at the core of our paper using a two-period model. In particular, following the literature, we assume that the first generation of users choose based on their private information and then post a review for their chosen option, while the second generation of users combines the first generation's ratings with their private information to form a posterior and make choices. In line with this, our theoretical analysis is divided into two distinct but tightly related parts: first, we analyze the properties of first period ratings as a function of the model's primitives. Then, we analyze their impact on the choices of second-period users.

We highlight three innovations: first, we model user heterogeneity in both choice and ratings. The former is driven by different precision in the users' private signals, while the latter takes a general form,

allowing us to study a large set of cases. Second, in line with empirical realism, we consider choices *between* products. Thus, we are interested in biases in *relative*, not *absolute*, ratings. Third, while a large literature on non-Bayesian social learning exists,[12] there is a lack of papers theoretically studying the impact of online reviews when users learn naïvely from them – that is, by simply comparing average scores. In recent work, Acemoglu et al. (2022) argue that this is an important direction for research.

2.2 Formal Set Up

There is a large number of users, equally divided into two generations. Users in each generation are divided into two types, experienced (E) and novices (N), in the proportions ψ and $1-\psi$, $\psi \in (0,\ 1)$. Each generation has the same composition of types. There are two products with qualities Q_H and Q_L, respectively, with $Q_H > Q_L$. Without loss of generality, we normalize the quality levels to $Q_H = 1$ and $Q_L = 0$. Thus, the quality gap $\Delta(Q) := Q_H - Q_L$, a key variable in our model, is also 1.[13]

Experienced users consume Mk products, while novices consume k, where $M > 1$. Without loss of generality, we can assume $k = 1$. Since we are ultimately interested in the impact of each user type on the information content of the platform, it is useful to think about the proportion of purchases made by each group of users, as opposed to the proportion of users themselves. The proportion of reviews is given by $M\psi/(M\psi + 1 - \psi)$ for E and $(1-\psi)/(M\psi + 1 - \psi)$ for N. For convenience, we denote these by ψ' and $1-\psi'$. Clearly, $\psi' > \psi$: E users are "overrepresented" on the platform due to their higher level of activity. Since ψ' is a sufficient statistic for our theoretical results, we will use ψ' instead of M and ψ separately in the rest of our theory section.

Each user observe a quality signal for each of the two products distributed as:[14]

$$s_j^i \sim \mathcal{N}(Q_i, 1/\tau_E), \quad i = H, L, \ j = E, N \tag{1}$$

That is, we assume each user observes an unbiased signal for each product, and that the signal's precision is (possibly) type-specific. Moreover, we assume signals are independent both across products and across users. τ_E and τ_N denote the precision of the E and N signals, respectively. In line with both our motivation

[12] See for instance Molavi, Tahbaz-Salehi and Jadbabaie (2018) and citations therein.

[13] We chose to develop our model with two quality levels and two user types for both clarity of exposition and mathematical tractability. Neither of these assumptions is critical.

[14] Our model adapts the well-studied normal-normal Bayesian framework introduced by Papanastasiou and Savva (2017) by adding novel elements that are crucial to our analysis: multiple products, two types of users, and endogeneity of ratings.

and empirical realism, we assume $\tau_E \geq \tau_N$: experience is key to observing (weakly) more (precise) signals of quality.[15,16] For example, experienced users might be more likely to read specialized blogs or magazines, talk to experienced peers, or pick up quality signals such as Oscar-winning directors or actors, as well as reviews by professional critics. (In Section 5.1.2, we show that this is indeed the case in both our IMDb and MovieLens data).

We start by characterizing the choice probabilities of the first generation of users. Given the absence of socially generated information, these depend solely on the private signals' precisions. Let p_E and p_N denote the probabilities that first-generation users E and N choose the high quality product. Given the signal structure for both E and N users, we have:

$$\begin{aligned} p_E &= prob\big(\mathcal{N}(1,2/\tau_E) \geq 0\big) = 1 - \Phi\left(\frac{-\sqrt{\tau_E}}{\sqrt{2}}\right), \\ p_N &= prob\big(\mathcal{N}(1,2/\tau_N) \geq 0\big) = 1 - \Phi\left(\frac{-\sqrt{\tau_N}}{\sqrt{2}}\right). \end{aligned} \tag{2}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution.

Clearly, $p_E,\ p_N > 1/2$ whenever $\tau_E, \tau_N > 0$. Furthermore, $\tau_E \geq \tau_N$ implies $p_E \geq p_N$.

We refer to the case $p_E > p_N$ as *choice heterogeneity* and $p_E = p_N$ as *choice homogeneity*. Given the one-to-one relationship between choice probabilities p_i and private information precision τ_i, $i = N, E$, all of our results can be stated using either. We use choice probabilities as they feed directly into our subsequent analysis.

After consuming a product, each first-generation user leaves a rating. (Given the very large number of users in our model, our results would be unchanged if each user left a rating only with a positive but fixed probability. We discuss the case of non-random self-selection into reviewing in Appendix A.1.) Without loss of generality, let ratings for each E (N) user to be given by, respectively,

$$\mathscr{R}_E(Q) = Q + \varepsilon_r, \quad \mathscr{R}_N(Q) = \alpha Q + \beta + \varepsilon_r, \ \ \alpha \geq 0.^{17} \tag{3}$$

[15] In this sense, our definition of experience aligns with the traditional one for expertise in economics, and particularly in the delegation literature – see for instance Crawford and Sobel (1982) and Krishna and Morgan (2001). In the context of online reviews, these two concepts have been used interchangeably (Nguyen et al., 2021). We opted to use experience as it allows us to draw a direct link with our empirical Section, in which we can proxy it with the number of reviews left by each user (which is also the approach taken by IMDb in defining its Top1000 users).

[16] From a mathematical point of view, this is equivalent to observing *more* (independent) signals of given precisions, whose precisions sum up to τ_E for E users and τ_N for N users.

[17] That is, given $\mathscr{R}_N(Q) = \alpha Q + \beta$ and $\mathscr{R}_E(Q) = \gamma Q + \delta$, the predictions are unchanged if we restrict attention to models in which $\gamma = 1$ and $\delta = 0$, as long as $\gamma \neq 0$ to begin with – a natural assumption, since its violation would imply that E's reviews would be independent of quality.

where $\varepsilon_r \sim \mathcal{N}(0, 1/\tau_r)$ represents an idiosyncratic shock, independent across both users (and, thus, users' categories) and products. Our assumption of a fixed τ_r across user categories, while not crucial for the analysis, is both analytically convenient and justified by our empirical findings.[18]

The ratings' precision τ_r is a key variable in our model, as it determines the role played by ratings in the choices of the second generation of users. When $\tau_r \to 0$, ratings are inconsequential and the second generation of users chooses solely based on their private information. When $\tau_r \to \infty$, ratings are the sole driver of second-period users' choices, independently of each user's prior beliefs.

While our data provide clear estimate of of α and, even more so, β in the context of movie ratings, in this Section we adopt an agnostic approach and spell out the potential consequences of user heterogeneity across the full range of α, β within the domain $[0,1] \times [-\bar{\beta}, \bar{\beta}]$, where $\bar{\beta} > 0$.[19] We categorize the potential scenarios into four intuitive and prominent categories:

- **Rating Homogeneity** ($\alpha = 1$, $\beta = 0$): E and N users leave the same ratings for each of the two products.

- **Experience Leads to Higher Diagnosticity** ($\alpha < 1$): N users' ratings are flatter with respect to quality. (In the limit case $\alpha = 0$, novice ratings are uninformative about quality.) This could be due, for instance, to an inability to appreciate either the faults of low quality products or the features of high quality ones, or both; it could also be due to an awareness of one's own inexperience, leading to more moderate evaluations.

- **Experience Leads to Higher Stringency** ($\alpha = 1$, $\beta > 0$): N users leave higher ratings for each quality level. That is, E and N users' ratings respond equally to quality *increases*. But N users' ratings show more leniency, regardless of quality. A behavioral microfoundation for this phenomenon in the case of $\tau_E > \tau_N$ is given by *expectation-based reference dependence* (Kőszegi and Rabin, 2006): Since E users choose better quality on average ($p_E Q_H + (1-p_E)Q_L > p_N Q_H + (1-p_N)Q_L$ whenever $\tau_E > \tau_N$), they form higher reference points – or standards – and are less satisfied for each quality level they experience.

- **Experience Leads to Lower Stringency** ($\alpha = 1$, $\beta < 0$): Contrary to what described above, it may

[18] Table 1 shows that the variances of ratings for users with different experience levels are statistically indistinguishable.

[19] Notice how we rule out two cases: $\alpha < 0$ and $\alpha > 1$. The first case implies a negative correlation between ratings and quality. The second corresponds to the scenario in which N users' ratings are overly sensitive to quality, which is implausible and clearly refuted in our empirical analysis. We also emphasize that, owing to the continuity in α, our results still hold as we transition from $\alpha = 1$ to $\alpha > 1$. Nevertheless, we opt to focus on the most pertinent cases to simplify the exposition.

be the case that E users leave higher ratings for each quality level. For example, E users may simply better equipped to recognize (and possibly derive greater enjoyment from) the positive features of the products they choose.

We start by assuming that the aggregate ratings for each product are given by the average of the individual ratings (See Section 2.4 for alternative aggregation rules.). Thus, given our specification, we have:[20]

$$\mathscr{R}_H = \mathscr{R}(1) = \frac{\psi' p_E \cdot Q_H + (1-\psi')p_N \cdot (\alpha Q_H + \beta)}{\psi' p_E + (1-\psi')p_N} = \frac{n_H^E + n_H^N(\alpha+\beta)}{n_H}, \tag{4}$$

$$\mathscr{R}_L = \mathscr{R}(0) = \frac{\psi'(1-p_E) \cdot Q_L + (1-\psi')(1-p_N) \cdot (\alpha Q_L + \beta)}{\psi'(1-p_E) + (1-\psi')(1-p_N)} = \frac{n_L^N \beta}{n_L}. \tag{5}$$

where n_i^j denotes the number (share) of users of type $j = N, E$ reviewing products of quality $i = L, H$, and thus $n_H = n_H^E + n_H^N = \psi' p_E + (1-\psi')p_N$, and $n_L = n_L^H + n_L^N = \psi'(1-p_E) + (1-\psi')(1-p_N)$.

In words, both $\mathscr{R}(1)$ and $\mathscr{R}(0)$ are weighted averages of N and E users' ratings. Crucially, products' qualities influence ratings in multiple ways: by directly changing $\mathscr{R}_E$ and $\mathscr{R}_N$; by changing the share of E users' ratings (whenever $p_E > p_N$); and by influencing the total number of ratings (given that $p_E, p_N \geq 1/2$).

2.3 Results

2.3.1 Properties of Ratings

We start by studying the properties of ratings left by the first generation of users. We start with two crucial definitions.

Definition 1 (**(Un)biased Ratings).** *Ratings are (relatively)* **unbiased** *whenever* $\Delta(\mathscr{R}) = \Delta(Q) = 1$*, and* **ratings compress qualities** *whenever* $\Delta(\mathscr{R}) < 1$.[21]

Notice that the two definitions above are robust to platform-wide ratings (dis)inflation: that is, they are independent on the values $\mathscr{R}_H$ and $\mathscr{R}_L$ *per se*. The definitions only speak to the *relative* magnitudes of these differences: for instance, $\Delta(\mathscr{R}) < 1$ is equivalent to $\mathscr{R}_H - 1 < \mathscr{R}_L(-0)$, or to the ratings of the higher quality product being *less* inflated (if at all) than those of the lower quality one.

[20] With a slight abuse of notation, we simply write $\mathscr{R}$ and not $\mathbb{E}(\mathscr{R})$ to refer to the average of aggregate ratings.

[21] The case $\Delta(\mathscr{R}) > 1$ is both impossible given our theoretical assumptions and strongly refuted by our empirical analysis, so we omit it in our main definition.

Definition 2 (**Ranking Reversal**). *A* **ranking reversal** *occurs whenever* $\Delta(\mathscr{R}) < 0$.

It is immediate to see that, given the definitions, a ranking reversal implies a compression. Whenever a rating compression occurs, we denote by $\mathscr{C} = 1 - \Delta(\mathscr{R})$ its degree. Clearly, $\mathscr{C} > 0$, and $\mathscr{C} > 1$ if and only if a ranking reversal occurs.

Our first two Lemmas highlight how the combination of the two forces at the core of our model – choice and rating heterogeneity – is a precondition for ratings to display pathological properties: shutting down (at least) one of these two forms of heterogeneity results in the absence of ranking reversals and, when $\alpha = 1$ and choices are homogeneous, ratings are unbiased for any β.

Lemma 1 (**Choice Homogeneity**). *Let* $\tau_E = \tau_N =: \tau$, *or equivalently* $p_E = p_N =: p$. *Then, for any admissible quadruple* $(\alpha, \beta, \psi', p)$, *ratings display no ranking reversals. Moreover, if* $\alpha = 1$, *for any triple* (β, ψ', p), *ratings are unbiased.*

Proof. The proof for this and all other results is relegated to Appendix A. □

Lemma 2 (**Rating Homogeneity**). *Let* $\alpha = 1$ *and* $\beta = 0$. *Then, for any admissible triple* (p_E, p_N, ψ'), *ratings are unbiased, and thus display no ranking reversals.*

Both Lemmas highlight straightforward facts. Lemma 1 shows that, when the share of E (or N) users is independent of a product's quality level, individual rating heterogeneity is not problematic, since aggregate ratings still represent an "apples-with-apples" comparison. A compression still occurs – for reasons orthogonal to user self-selection – whenever N users' ratings flatten qualities ($\alpha < 1$), but no ranking reversal can occur. Lemma 2 formalizes how, when E and N users rate equally, their self-selection patterns – as dictated by p_E and p_N – are inconsequential in determining the average of ratings.

What conditions on α, β, τ_E and τ_N imply the presence of ratings compressions? The following Proposition formalizes necessary and sufficient conditions for rating compressions to arise, as well as all their relevant comparative statics in the model parameters.

Proposition 1 (**Rating Compression**).

- *Let* $\alpha = 1$ *and* $p_E > p_N$. *Then, a* **rating compression** *occurs if and only if* $\beta > 0$, *for every* ψ', p_E, *and* p_N. *Moreover, the degree of compression* $\mathscr{C}$ *is increasing in* β *and* p_E, *decreasing in* p_N, *and inverse U-shaped in* ψ'.

- *Let $\alpha < 1$ and $p_E > p_N$. Then, there exists a $\beta^* = \beta^*_{RC}(\alpha, \psi', p_E, p_N) < 0$ such that a rating compression occurs whenever $\beta > \beta^*_{RC}$. Moreover, β^*_{RC} is increasing in α and p_E, and decreasing in p_N. Lastly, the degree of compression $\mathscr{C}$ is decreasing in α.*

Proposition 1 has some interesting implications. First, intuitively, when E and N users' ratings respond to quality increases equally ($\alpha = 1$), a rating compression is equivalent to E users having higher stringency ($\beta > 0$). However, when N users respond less to quality increases ($\alpha < 1$), $\beta \geq 0$ is no longer necessary for a rating compression. While $\beta < 0$ implies that E users are less stringent, this is (more than) compensated by the flat slope of N users' ratings whenever α is small.

As we will see in Section 2.3.2, a rating compression can potentially be damaging to second generation users' choices even absent ranking reversals. This is because, while potentially imprecise, users' private information is unbiased: on average, $\mathbb{E}(s_j^H - s_j^L) = 1$ for $j = E, N$. The introduction of compressed ratings increases each agent's beliefs' precision, but at the cost of introducing bias ($\Delta(\mathscr{R}) < 1$), thus potentially increasing the fraction of users that form a higher posterior belief for the low quality product than the high quality one (Theorem 4).

The next Proposition takes one step further, by highlighting that there exists a β such that ranking reversals will occur.

Proposition 2 (Ranking Reversals).

For every quadruple $(\alpha, \psi', p_E, p_N)$, there exists a $\beta^*_{RR} = \beta^*_{RR}(\alpha, \psi', p_E, p_N)$ such that, if $\beta \geq \beta^*_{RR}$, ratings display a **ranking reversal***: $\mathscr{R}(0) > \mathscr{R}(1)$. Moreover, $\beta^*_{RR} > 0$, and is increasing in α and p_N, decreasing in p_E, and U-shaped in ψ'. Last, for any quadruple $(\alpha, \psi', p_E, p_N)$, $\beta^*_{RR} > \beta^*_{RC}$.*

Proposition 2 formalizes a stark result, namely, the existence of a positive β^*_{RR} such that ranking reversals occurs whenever $\beta \geq \beta^*_{RR}$, independent on all the other model fundamentals: private information (and thus choice probabilities) of both E and N users (p_E, p_N), and *a fortiori* the (positive) degree of their heterogeneity; the slope of the mapping of quality into ratings by N users (α); and the proportion of overall purchases, and thus ratings, left by E users (ψ').

To grasp some intuition, notice that the aggregate ratings of each of the two products are a weighted average between its ratings from E and N users. The key observation is that, as β grows large, N's ratings of the lower quality products eventually become higher than E's ratings of the high quality product. This leads to a ranking reversal whenever choice heterogeneity is strong enough: if the low quality product is rated by

predominantly N users, and the high quality one by predominantly E users, the former will obtain a higher average rating.

The comparative statics of β^*_{RR} with respect to $\alpha, \psi', \tau_E, and \tau_N$ illuminate the logic behind the proposition. Intuitively, we would expect that more choice heterogeneity leads to more ranking reversals, whereas larger quality differences make this possibility less likely. Indeed, this is confirmed in our comparative statistics: an increase in p_E, all else fixed, increases choice heterogeneity, whereas one in p_N decreases it. Moreover, a higher α makes N's ratings more responsive to quality, thus increasing $\Delta(\mathscr{R})$ and decreasing the possibility of a choice reversal.

In other words, the degrees of choice and rating heterogeneity (which we have shown to jointly be needed for the occurrence of compressions, and thus reversals – see Lemma 1 and 2) are complementary in causing ranking reversals. When choices are relatively similar (small $p_E - p_N$), a large degree of rating heterogeneity is needed; conversely, when $p_E - p_N$ grows large, even a small amount of rating heterogeneity suffices.

The less intuitive comparative statics are the ones in ψ' – a feature shared by Propositions 1 and 2. Where does the non-monotonicity come from, especially in light of the linear mapping of qualities into ratings? Put differently, why is it that an increase in ψ' increases the probability of a ranking reversal when ψ' is small, while decreasing it when ψ' is high? The reason for this lies in the amount of user heterogeneity on the platform. When ψ' is high – either because E users are disproportionately more active or because their baseline proportion, ψ, is high – a further increase makes the platform more homogeneous, thus mitigating the inefficiencies stemming from user heterogeneity. Conversely, when ψ' is small, an increase in ψ' increases heterogeneity. The following, simple example succinctly illustrates the idea.

Example 1 (U-Shaped Effects of User Heterogeneity)**.** *Let $p_E = 0.9$, $p_N = 0.5$, and $\alpha = 1$. Then, upon multiplying both $\mathscr{R}(1)$ and $\mathscr{R}(0)$ by $10(1-\psi')$, and denoting $\psi'' = \psi'/(1-\psi')$ for expositional simplicity, we have:*

$$\mathscr{R}(1) = \frac{9\psi'' + 5(1+\beta)}{9\psi'' + 5}, \;\; \mathscr{R}(0) = \frac{5\beta}{\psi'' + 5}.$$

Straightforward algebra[22] *shows that, for a fixed β, $\Delta(\mathscr{R})$ is decreasing when $\psi'' < 5/3$ and increasing when $\psi'' > 5/3$. This cutoff corresponds to $\psi' = 5/8$, or approximately 71%. Moreover, $\mathscr{C}(0.71) = 0.13$. Put differently: the rating compression is most severe when $\psi' = 0.71$. When E users are responsible for*

[22] See Appendix A.

more than 71% *of ratings on the platform, the bias would decrease with more E users. If, however, E users account for less than* 71% *of the ratings, the opposite is true: aggregate ratings would become more compressed with more E users.*

Let $p_E = 0.8$ *now, and everything else unchanged. Then, it is easy to show that* $\Delta(\mathscr{R})$ *is increasing when* $\psi' \geq 0.8$*. Moreover,* $\mathscr{C}(0.8) = 0.13$*. The smaller degree of rating compression –* 0.13 *vs* 0.25 *– is a natural consequence of the lower degree of choice heterogeneity.*

2.3.2 Period Two: (Mis)Learning from (Biased) Ratings

A second, identical generation of users arrive in the following period. Each second generation user observes both a private signal of quality (whose distribution is exactly the same of their predecessors', described in Equation 1), and the ratings posted in period 1.

Recall that n_i, $i = H, L$ indicates the number of ratings for each of the two products. Given n_i and $\mathscr{R}(Q_i)$, as well as ratings' informativeness τ_r, each user observes a signal coming from reviews,

$$s_i \sim \mathscr{N}(\mathscr{R}(Q_i), 1/n_i\tau_r), \quad i = H, L.$$

Given prior beliefs $s_{ij} \sim \mathscr{N}(Q_i, 1/\tau_j)$, users of type $j = E, N$ form a posterior equal to:

$$\tilde{Q}_{ij} \sim \mathscr{N}\left(\frac{\tau_j Q_i + n_i\tau_r\mathscr{R}_i}{\tau_j + n_i\tau_r}, \frac{1}{\tau_j + n_i\tau_r}\right), \quad i = H, L. \tag{6}$$

Notice that $\tau_r = 0$ corresponds to the case of no (impact of) reviews, and thus of second generation users choosing according to their prior beliefs. Conversely, $\tau_r \to \infty$ corresponds to the case of infinitely precise reviews, which render private information irrelevant, thus making second period users' choices homogeneous both within and between user types.

Equation 6 highlights an important fact: both the mean and the precision of the posterior depend on both products' quality level i and on users' type j. On the products' side, the dependence of the mean on the underlying quality of each product is relatively straightforward, as well as, of course, desirable. The direct dependence on Q_i is moderated by the relative precision of private information and first generation reviews. On the users' side, we note that in the non-trivial cases $\tau_r < \infty$, ratings have differential impacts on users' belief updating: N users rely on ratings more than E whenever $\tau_E > \tau_N$.

What is the impact of ratings on second generation users' choices? Equation 6 implies that posterior

beliefs for the quality gap between the two products for consumers $j = E, N$ are given by

$$\tilde{\Delta}(Q)_j \sim \mathcal{N}\left(\frac{\tau_j + n_H \tau_r \mathscr{R}_H}{\tau_j + n_H \tau_r} - \frac{n_L \tau_r \mathscr{R}_L}{\tau_j + n_L \tau_r},\ \frac{1}{\tau_j + n_H \tau_r} + \frac{1}{\tau_j + n_L \tau_r}\right). \tag{7}$$

Thus, a user of type $j = E, N$ selects the higher quality product if and only if $\tilde{\Delta}(Q)_j > 0$.

2.3.3 Benchmark: Learning from Unbiased Ratings

When ratings reflect qualities, it is immediate to see that Equation 7 reduces to:

$$\tilde{\Delta}(Q)_j \sim \mathcal{N}\left(1,\ \frac{1}{\tau_j + n_H \tau_r} + \frac{1}{\tau_j + n_L \tau_r}\right). \tag{8}$$

This simple fact implies a natural corollary: when unbiased, ratings improve consumer welfare for both E and N users.

Proposition 3 (Learning from Unbiased Ratings)**.** *When $\mathscr{R}_H = 1$ and $\mathscr{R}_L = 0$, ratings improve the choices of second generation of consumers. Moreover, the improvement is increasing in τ_r.*

2.3.4 (Mis)Learning from Compressed Ratings

We start from the case of a rating compressions that do not result in a ranking reversal: $\mathscr{C} \in (0, 1)$. In this scenario, what is the impact of ratings on 2nd generation users' choices and welfare? The answer is subtle, as it depends on both the degree of compression $\mathscr{C}$ and the ratings' precision, τ_r. We start with two Lemmas, highlighting sufficient conditions for ratings to improve welfare.

Lemma 3. *When $\tau_r \to \infty$, ratings improve the welfare of of both E and N users, independent on their degree of compression $\mathscr{C} \in (0, 1)$.*

Lemma 3 formalizes an intuitive fact: when ratings become very precise, it suffices for them to rank the two products correctly. This is because, when $\tau_r \to \infty$, Equation 7 becomes:

$$\tilde{\Delta}(Q)_j \sim \mathcal{N}\left(\frac{\tau_j + n_H \tau_r \mathscr{R}_H}{\tau_j + n_H \tau_r} - \frac{n_L \tau_r \mathscr{R}_L}{\tau_j + n_L \tau_r},\ \frac{1}{\tau_j + n_H \tau_r} + \frac{1}{\tau_j + n_L \tau_r}\right) \xrightarrow{p} \mathscr{R}_H - \mathscr{R}_L =: 1 - \mathscr{C} > 0. \tag{9}$$

Thus, despite underestimating quality differences, all second period users end up ranking the two options correctly, thus picking the higher quality one.

Our next lemma is concerned with how the severity of the compression itself impacts welfare.

Lemma 4. *There exists a $\mathscr{C}^* = \mathscr{C}^*(\tau_E, \tau_N, \tau_r)$ such that, if $\mathscr{C} < \mathscr{C}^*$, ratings improve the welfare of both E and N users, independent on τ_r. Moreover, for every τ_N, τ_E and τ_r, $\mathscr{C}^* \in (0,1)$.*

Lemma 4 proves the existent of a non-trivial cutoff $\mathscr{C}^*$ such that, if the compressions is less severe than $\mathscr{C}^*$, ratings increase welfare despite being biased.

What happens when both of these conditions are violated? We have the following:

Proposition 4 ((Mis)Learning from Compressed Ratings). *Fix τ_E and τ_N. Then, there exists a $\overline{\mathscr{C}} < 1$ such that, if $\mathscr{C} > \overline{\mathscr{C}}$, ratings hurt the welfare of both E and N users for intermediate values of τ_r.*

Proposition 4 highlights a crucial result: compressed ratings can hurt the welfare of all users, even absent ranking reversals, and when all users agree on the products' relative qualities. This is because ratings add bias to consumers' private information, which is noisy but unbiased. Based on the interplay between relative ratings' bias and precision, the information distortion caused by ratings compressions can strictly decrease the probability that both E and N users choose the higher quality product.

Combined, Lemmas 3 and 4 and Proposition 4 imply a non-monotonic effect of ratings' precision on both E and N users' welfare whenever $\mathscr{C} > \mathscr{C}^*$: when ratings precision tends to 0, second generation users choose according to their prior information, just like their predecessors, thus choosing the high quality option with probability $p_j, j = E, N$. As ratings' precision grow from 0 to a cutoff τ_r^*, this probability – and therefore users' welfare – decreases to $\underline{p}_j \in (0, p_j)$. It then increases to its maximum of 1 as τ_r grows large.

2.3.5 (Mis)Learning with Ranking Reversals

We now turn our attention to the more pathological case highlighted in Proposition 2: ranking reversals, or $\Delta(\mathscr{R}) < 0$ (or equivalently, $\mathscr{C} > 1$). Whenever a ranking reversal occurs, we have the following:

Proposition 5 ((Mis)Learning with Ranking Reversals). *Let $\mathscr{C} > 1$. Then, ratings worsen all users' choices, the more so the higher τ_r.*

Proposition 5 formalizes a straightforward fact. Unlike in the case $\Delta(\mathscr{R}) \geq 0$, in which ratings contained some information about relative qualities, here ratings lead consumer astray. The stronger their impact on users' posteriors, the larger this effect becomes. Mathematically, when τ_r grows large, the opposite of what described in Equation 9 happens, since $1 - \mathscr{C} < 0$. Thus, second generation users become very confident, but wrong, about the relative ranking of the two options, minimizing their welfare.

2.4 Platform Design Implications

What can platforms do to minimize mislearning from ratings and improve their consumers' welfare?

So far, we have assumed that the platform simply takes an unweighted average of each individual rating, irrespective of users' experience level. Several platforms, however, have adopted slightly more sophisticated aggregation rules, in an attempt to improve on their users' collective wisdom. Among these, a prominent one is a "seniority rule" that assigns extra weight to the opinions of more experienced members. This can help whenever these opinions are more reflective of actual product quality than novices' are. On the flip side, it could worsen the aggregation bias we highlight by adding an extra penalty on products that are rated by a higher share of E users.

Our next result – which follows from the comparative statics of Propositions 1 and 2 – highlights necessary and sufficient conditions for this policy to backfire.

Corollary 1 (Overweighting Experienced Users' Opinions)**.** *Let the platform overweight E's ratings by a factor* $\gamma > 1$*:*

$$\mathscr{R} = \frac{\gamma \cdot \psi' p_E \mathscr{R}_E + (1-\psi') p_N \mathscr{R}_N}{\gamma \cdot \psi' p_E + (1-\psi') p_N}. \tag{10}$$

Then, for each $\gamma > 1$*, ratings will become less (more) compressed whenever* $\psi' > (<) \psi^*(\alpha, \beta, p_E, p_N)$*, where* ψ^* *denotes the unique minimizer of* $\Delta(\mathscr{R})$ *defined in Proposition 1.*

Where does this dichotomy come from? Intuititively, the platform faces a trade-off between the diagnosticity of individual ratings (which is higher for E users whenever $\alpha < 1$) and their aggregation across users types. The latter becomes more severe whenever the shares of the two user types' contributions on the platform are fairly balanced – the more so the more ratings and choices differ between users types.

Overweighting the more experienced users' ratings is equivalent to increasing the share of their contributions from ψ' to $\gamma\psi'/\gamma\psi' + (1-\psi') > \psi'$, and therefore exacerbates the problem by making the *"apples-with-oranges"* comparison more salient.

If the platform is aware of the underlying rating heterogeneity, can it not simply rescale individual ratings? In Section 6, we show how the platform can shut down rating heterogeneity by computing the leniency of each user and then correcting for it. Of course, changing each user's leniency in turns changes aggregate ratings – feeding back into leniences. Thus, this solution requires a recursive approach.[23]

[23] Our algorithm is slightly more complex than its theoretical counterpart, reflecting the granularity of our data: on MovieLens, we are able to track individual users, and thus account for individual differences in stringency, as well as the evolution of individual stringency over time.

3 From Theory to Data

Our model presents a series of results regarding how user heterogeneity can bias the relative ratings displayed by online platforms. In particular, it shows how the presence (and degree) of ratings compressions depends on a combination of choice and rating heterogeneity.

We now turn to studying these issues empirically: do we observe experience-driven choice heterogeneity and/or rating heterogeneity? If so, of what kind? Informed by our model, what can we do to improve on platform's design? Our approach is two-fold. First, we directly test for both choice and rating heterogeneity (Section 5). While the latter is fairly straightforward (we can simply compare, product by product, ratings for different user types, see Section 5.2), the former requires us to use external quality information – an issue we tackle by complementing our data with products' awards and other quality signals (Section 5.1).

We then propose a methodology to de-bias the ratings (Section 6), a complementary and almost entirely model-free approach. Without making any assumptions on users' rating and choice processes, we develop a simple, recursive algorithm to rescale the ratings. This not only leads to substantial changes in ratings (Section 6.3), but also provides additional validation for both our model's assumptions (Section 2) and our regression estimates (Section 5).

4 Dataset

4.1 MovieLens and IMDb

Our dataset merges data from two online movie recommendation systems, MovieLens and IMDb, encompassing movie information, user ratings, and user demographics. MovieLens is a web-based recommender system with a small, but dedicated community of users. It is a project of GroupLens Research at the University of Minnesota, and it allows users to rate movies and receive recommendations. For our analysis, we use the "MovieLens 25M" Dataset, containing 25 million individual ratings from January 1995 to November 2019.[24,25] IMDb, in contrast, is the largest online platform for user-generated movie content. It was one of Amazon's first acquisitions in 1998. As of January 2024, it is the 4th most visited website in the US in the streaming & online TV category (behind YouTube, Max, and Netflix), and the 65th most visited site

[24]For more information about MovieLens, see Harper and Konstan (2015) and `https://files.grouplens.org/datasets/MovieLens/ml-25m-README.html`.

[25]For recent studies using MovieLens data, see Aridor, Gonçalves, Kluver, Kong and Konstan (2022) and discussion therein.

globally in any category.[26] For the purposes of our analysis, in addition to its scale, it offers a broader range of aggregate data on movies and user demographics than MovieLens, but lacks comprehensive user rating histories.

4.2 Dataset Construction

Our analysis focuses on 9,426 movies that meet two criteria: they have been rated by at least 30 users in our MovieLens dataset and were produced after 1994. For each of these movies, we also obtain detailed information from IMDb. The combination of MovieLens and IMDb data is key, as the two datasets complement each other nicely. Moreover, we ensure the robustness of our results by validating them across both platforms. MovieLens's individual user ratings help analyze rating history and stringency, crucial for our de-biasing algorithm (see Section 6). IMDb, while limited to aggregate ratings, offers comprehensive movie attributes and user demographics. To identify experienced users, we use IMDb information about the Top1000 users. These are the 1,000 users "who have voted for the most titles on the webpage." The identities of these users and the number of movies each of them has rated are not disclosed by IMDb. This ensures that users are unaware of their role, ruling out socially-driven explanations behind their rating behavior (Jacobsen, 2015). IMDb stopped disclosing information about the Top1000 users' ratings since 2023.[27] However, we managed to scrape IMDb movies' webpages in 2022, providing us with information about both the number and the average of ratings from Top1000 users for each movie at the time of scraping. Moreover, we have data on IMDb user age groups (18-29, 30-44, over 45), gender distribution, and the proportion of U.S. users per movie. Finally, we augment our dataset by including additional information on each movie's release year, runtime, genre, Academy Award nominations and wins, director and main cast, average ratings on both platforms, total user ratings, Metacritic reviews, and aggregated Metacritic scores.[28]

4.3 Descriptive Statistics

Table 1 presents the descriptive statistics for the 9,426 movies in our dataset. The statistics cover movie genres, runtimes, production years, and characteristics of the users who rated them. Approximately 70% of the movies fall into the genres of action, comedy, or drama. Receiving nominations or awards from the

[26] For this and more information about IMDb, see: https://www.similarweb.com/website/imdb.com/.

[27] Appendix Figure B1 displays a snapshot of the webpage containing the rating information for the film "Toy Story" before IMDb ceased to disclose details about the Top 1000 users' ratings.

[28] For information on Metacritic's review aggregation, visit https://www.metacritic.com/about-metascores.

Table 1. Summary Statistics: Movies' Characteristics, Ratings, and Audience

	Mean	SD	N	Min	Max
movie characteristics					
Year of Production	2007	6.81	9426	1995	2019
Movie Runtime	1.735	.482	9426	.0333	10.48
Genre: Action (%)	17	.	9426	.	.
Genre: Comedy (%)	26	.	9426	.	.
Genre: Drama (%)	26	.	9426	.	.
nominations, awards, and critic reviews					
Academy Nominated (%)	10	.	9426	.	.
Academy Awarded (%)	3	.	9426	.	.
Academy Nominated or Awarded: Director (%)	15	.	9426	.	.
Academy Nominated or Awarded: Star (%)	20	.	9426	.	.
Academy Nominations	.171	.672	9426	0	11
Academy Awards	.054	.409	9426	0	11
Director Previous Academy Nominations	.426	1.64	9426	0	24
Main Star Previous Academy Nominations	.509	1.61	9426	0	63
Nominations or Awards (No Academy)	16.4	37.8	9426	0	573
Have Critics Reviews: $Meta_i$ (%)	72	.	9426	.	.
Have Positive Critics Reviews: $Meta_i^{>60}$ (%)	32	.	9426	.	.
movie ratings					
IMDb Ratings: $\bar{r}_i^{IMDb}$	6.51	.979	9426	1.4	9.5
IMDb Number of Ratings: $n_i^{IMDb}(\times 1000)$	68.95	148	9426	.05	2588
IMDb Top1000 Ratings: $\bar{r}_i^{Top1000}$	5.95	.864	9426	1	9
IMDb Number of Top1000 Ratings: $n_i^{Top1000}$	260.3	191	9426	2	928
MovieLens Ratings: $\bar{r}_i^{Movielens}$	3.24	.459	9426	.8548	4.483
MovieLens Number of Ratings: $n_i^{Movielens}(\times 1000)$	1.571	4.36	9426	.031	72.67
MovieLens Top 1% Ratings: $\bar{r}_i^{1\%}$	3.03	.484	9426	.7	4.292
MovieLens Top 1% Number of Ratings: $n_i^{1\%}$	184.3	262	9426	1	1578
movie audience					
Share 18-29 Users: p_i^{18-29} (%)	13	.	9426	.	.
Share 30-44 Users: p_i^{30-44} (%)	60	.	9426	.	.
Share Over45 Users: p_i^{45} (%)	26	.	9426	.	.
Share Female Users: p_i^{female} (%)	21	.	9426	.	.
Share US Users: p_i^{US} (%)	30	.	9426	.	.

Note: The table includes all 9,426 scraped movies and presents movies' characteristics; average ratings and number of ratings by all reviewers and Top1000 on IMDb and MovieLens; and the profile of movies' audience on IMDb.

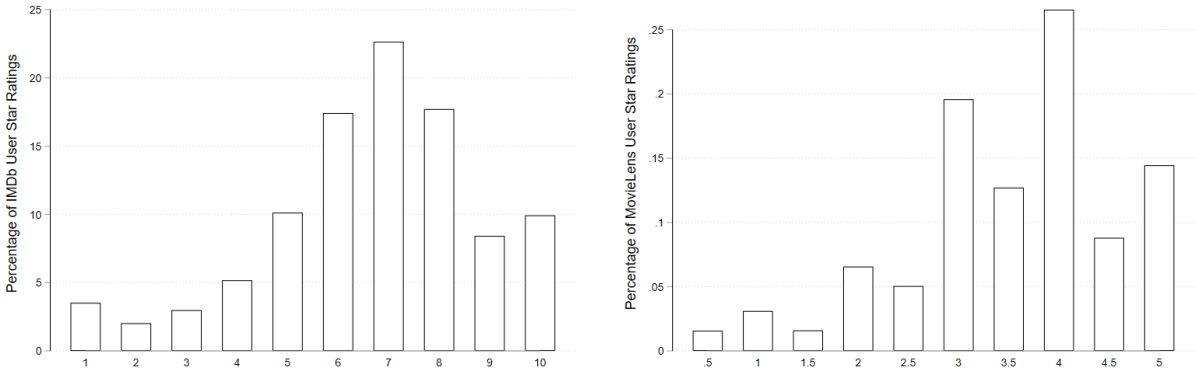
Academy of Motion Picture Arts and Sciences, commonly known as the Oscars, is relatively uncommon. Only 10% of the films in the sample have received at least one Oscar nomination, and 3% have been awarded an Oscar. On the other hand, 15% of the movies are directed by directors who have received at least one nomination in the past, and 20% feature actors or actresses who have received at least one nomination. This suggests a substantial presence of directors and actors with prior recognition in the industry. On average, the movies have received 0.17 Oscar nominations and 0.05 Oscar wins. However, movies often receive nominations and awards from other film festivals and industry accolades. On average, the movies in the dataset receive a total of 16 nominations or awards from sources other than the Oscars. Moreover, over 70% of the films have been reviewed by at least one film critic chosen by Metacritic, and a third of movies in our dataset have received positive evaluations, indicated by a Metascore exceeding 60 out of 100.

The distributions of ratings on IMDb and MovieLens are comparable. On IMDb, the average rating is 6.5 stars out of 10, with an almost one-star standard deviation. On MovieLens, the average rating is 3.2 stars out of 5, with a half-star standard deviation. Additionally, the average movie in our sample has been rated by several thousand users on both platforms.

IMDb's Top1000 users tend to post lower ratings compared to Non-Top1000 users. This finding represents the first aggregate evidence of a significant difference in rating patterns between Top1000 users and other IMDb users. However, the dispersion of ratings among the Top1000 users is similar in size to that of Non-Top1000. Similarly, on MovieLens, the top percentile of users, based on the number of ratings they have submitted, also tend to post lower ratings compared to the average MovieLens user. Overall, these observations suggest that there are differences in rating behavior between different user groups on both platforms, but the dispersion of ratings within these groups remains relatively consistent. We delve into this in much greater detail in Section 5.2, in which we compare ratings for the same movies given by both Top1000 and non-Top1000 IMDb users. Lastly, the demographic breakdown of IMDb users reveals that 60% of them fall within the age range of 30 to 44 years old. The majority of IMDb users are male. Around 30% of IMDb users are located in the US.

4.4 Selection into Reviewing

Before moving on to Section 5, in which we investigate the presence and nature of both choice and rating heterogeneity, we stress that we can only observe users' ratings, not choices. This would affect the interpretation of our results when the probability of rating a product, conditional on choice, is not random, and



(a) IMDb Star Ratings Percentage

(b) MovieLens Star Ratings Percentage

Figure 1. Distribution of IMDb and MovieLens Star Ratings Proportions

instead depends on characteristics of the product (or user).

Previous research has documented an *extremity* bias (Dellarocas and Narayan (2006); Schoenmueller et al. (2020); Karaman (2021); Brandes et al. (2022)): users tend to rate products only when they are very satisfied or dissatisfied with them. This bias is evident in the rating distributions of various online products, which often exhibit a "j-shaped distribution" of ratings with a majority of high scores, very few medium scores, and some low scores (Hu, Pavlou and Zhang, 2009; Hu, Pavlou and Zhang, 2017). However, Schoenmueller et al. (2020) show that extremity bias does not uniformly affect all platforms, and that "information platforms" with a larger scale for ratings, like MovieLens and IMDb, exhibit less polarity imbalance.

In this Section, we present empirical evidence indicating that the issue of self-selection at the rating stage may be less significant on the platforms we study, compared to other contexts. Figure 1 shows the distribution of star rating percentages for IMDb and MovieLens. Notably, we do not observe a j-shaped (or even u-shaped) distribution of ratings on either platform. Instead, the ratings tend to follow a roughly normal distribution (which also supports our theoretical modeling choices). If anything, there is a slight right skew in the ratings, consistent with a consumption-weighted distribution of movie quality.

Nonetheless, the lack of j-shaped ratings at the platform level does not necessarily mean that a bias is absent for some specific movies.[29] To address this concern, we identified movies with j-shaped ratings distributions on both platforms. Specifically, we say a movie's rating distribution is "j-shaped" if it received both more negative scores (1, 2, 3 stars for IMDb, and 1, 2 for MovieLens) and more positive scores (7, 8, 9, 10 stars for IMDb, and 4, and 5 for MovieLens) than medium scores (4, 5, 6 stars for IMDb, and 3 for

[29] We note that, in principle, this could be the case for highly polarizing movies even absent any selection into reviewing.

MovieLens).[30] Even applying this very conservative criterion, we only remove 60 movies from the sample, which represents 0.06% of the total. In the following Sections, we replicate our analysis excluding these movies with j-shaped or u-shaped ratings, and all results are confirmed.[31]

5 Results

5.1 Choice Heterogeneity

In Section 2, Experienced users differ from Novices since they observe more precise signals of quality, leading them to be more likely to choose high-quality products. As an empirical analog for this assumption, we expect IMDb Top1000 users to be more likely to rate high-quality movies. The main estimating equation to study this assumption is the following:

$$\ln(\pi_i^{Top1000}) = \alpha + \beta_1 q_i + \beta_2 \mathbf{X}_i + \varepsilon_i. \tag{11}$$

Here, $\ln(\pi_i^{Top1000})$ represents the natural logarithm of the ratio between the number of ratings posted by Top1000 and Non-Top1000 users on IMDb for movie i.[32] q_i is the unobserved quality for movie i, $\mathbf{X}_i$ is a set of controls accounting for movie characteristics, popularity, and audience, and ε_i denotes the error term.

Movie quality is not perfectly observable. In our analysis, we use four measures as proxies for quality that are independent of platforms' feedback. These measures are: the number of Academy Award nominations and wins received by each movie; the historical record of Academy Award nominations and wins by the movie's director or main star in the years preceding the movie's production; and the number of nominations and awards received by the movies from a wide range of film festivals and industry awards. Finally, we consider whether the movies have received a positive Metascore.

In Table 2, we use the number of Academy nominations and awards received by each movie as a proxy

[30] Our definition of a j-shaped distribution may be considered generous, as it includes movies with a u-shaped distribution as well. However, the primary objective of this exercise is to exclude any movies that could potentially exhibit even a slight resemblance to extremity bias in their ratings.

[31] A complementary approach would be to examine the distribution of ratings at the user level and determine whether some users have a tendency to post u-shaped or j-shaped ratings. While this analysis is not possible using the IMDb dataset, it can be performed using the full MovieLens 25M dataset, which includes individual ratings from MovieLens users. Using the same definition of j-shaped distributions used for movie rating distributions, we find that 98.5% of users do not have j-shaped distributions. Furthermore, the difference in the total number of posted ratings between those with and without j-shaped rating distributions is not statistically significant.

[32] Appendix Figure B2 shows the distribution of the number of Top1000 ratings for all movies in our sample. No movies are rated by all 1000 Top1000 users. Accordingly, issues related to censoring bias are not relevant to our analysis.

Table 2. Top1000 Users Are More Likely to Rate Movies with Nominations and Awards

	(1)	(2)	(3)	(4)	(5)	(6)
Academy Nominations	0.05***	0.04***	0.04***			
	(0.01)	(0.01)	(0.01)			
Academy Awards				0.06***	0.08***	0.08***
				(0.02)	(0.01)	(0.01)
Movie Runtime	-0.26***	-0.26***	-0.22***	-0.25***	-0.26***	-0.22***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$n_i^{critics}(\times 1000)$	-4.59***	-3.60***	-3.63***	-4.55***	-3.58***	-3.61***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
$n_i^{Movielens}(\times 1000)$	-0.06***	-0.08***	-0.08***	-0.06***	-0.08***	-0.08***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
p_i^{female}			-0.66***			-0.66***
			(0.05)			(0.05)
p_i^{US}			1.25***			1.25***
			(0.04)			(0.04)
Constant	-3.60***	-3.67***	-3.98***	-3.60***	-3.67***	-3.98***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.599	0.655	0.687	0.599	0.656	0.687
N	9,426	9,426	9,426	9,426	9,426	9,426

Note: The outcome variable is the log of the ratio between the number of ratings posted by Top1000 users and the number of ratings posted by non-Top1000 users. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

for quality. We run three different specifications, each controlling for movie characteristics, and audience, to avoid potential confounding effects on our estimates. The results indicate a positive correlation between receiving Academy nominations or awards and a higher proportion of Top1000 users posting ratings for a movie. Specifically, an additional Academy nomination is associated with an increase of almost 5% in Top1000 user ratings relative to Non-Top1000 user ratings, and an additional Academy award is associated with an increase of more than 6%. This finding is consistent with our model's assumption that experience is instrumental in selecting higher-quality products. However, it is also surprising given that winning an Oscar is often seen as a popularity boost for movies. Given their increased popularity, Oscar-nominated or Oscar-winning films should theoretically attract a broader audience, including those who are less experienced and less likely to watch movies. For this specific reason, we control for other factors that correlate with a movie's overall popularity. These factors include the number of Metacritic critics ($n_i^{critics}$) who have reviewed the movie and the number of users who have rated the movie on MovieLens ($n_i^{MovieLens}$). The finding that more experienced users are more likely to rate movies with Academy nominations and awards is highly robust

across different model specifications.

In Appendix Table B1, we introduce two categorical variables to indicate whether a movie has not received, received at least one, or more than one Academy nomination or award. Using the same specifications in Table 2, the proportion of ratings by Top1000 users is significantly higher for those movies that receive more than one Academy nomination or award. This outcome is consistent with the notion that movie quality is more strongly associated with films that receive multiple nominations or awards rather than just one. In Appendix Table B2, we replicate the specifications presented in Table 2, this time focusing on the subset of movies that do not exhibit a "j-shaped" rating distribution on both IMDb and MovieLens platforms. All the results from our analysis consistently reaffirm the positive and statistically significant relationship between movie quality and the proportion of Top1000 users who rate these movies.

Up to this point, we have used the logarithm of the ratio of the number of reviews posted by Top1000 users to the number of reviews posted by non-Top1000 users. We prefer this specification because it allows us to interpret the coefficients of our estimations in terms of percentage changes. This aspect is particularly important since the Top1000 users represent a very small minority of users, and for several films the ratio is close to zero. Nevertheless, the sign of the coefficient does not depend on the logarithmic functional form. The results are also robust to using the raw ratio between the number of ratings posted by Top1000 users and the number of ratings posted by non-Top1000 users, as shown in Appendix Table B3.

We now turn to MovieLens to check whether the same pattern holds true. To this end, we first identify the top 1% of MovieLens users who have contributed the highest number of ratings on the platform. We then calculate the logarithm of the ratio between the number of ratings posted by the top 1% of MovieLens users and the number of ratings posted by the remaining 99% of MovieLens users ($\ln(\pi_i^{1\%})$). This serves as our first proxy. As our second proxy, we consider IMDb information on the proportion of movie ratings from different age groups. We believe that age might be plausibly correlated with experience, so we calculate the logarithm of the ratio between the number of ratings posted by users over 45 years old and the number of ratings posted by under 29 and over 18 years old ($\ln(\pi_i^{over45})$). Appendix Table B4 presents the specifications shown in Columns (3) and (6) of Table 2 for these two proxies. The top 1% of users are more likely to rate movies that have received Academy Awards. However, the receipt of Academy nominations does not seem to significantly impact the proportion of ratings from MovieLens most experienced users. In contrast, movies nominated for or awarded by the Academy are more likely to be rated by users over the age of 45.

In the remainder of this Section, we analyze two potential limitations with our focus on the number of

Oscar nominations and wins to proxy movie quality. The first issue concerns the timing of these awards, as reviewers may have seen and rated the films before or after they were nominated or won Oscars and the popularity boost of the awards could confound our estimates. The second limitation is that while the Oscars tend to recognize high-quality films, they represent only a small fraction of all films. In addition, experienced users should be able to discern the quality of movies not only by watching movies that (will) win or are (will be) nominated for the most prestigious awards, but also by recognizing other high-quality movies that may have less prominent indicators of excellence.

5.1.1 The Academy Awards Timing

The IMDb dataset does not include the dates when Top1000 and non-Top1000 users posted their ratings. Hence, we are unable to directly address the confounding element related to the timing of awards and nominations by restricting our analysis to users who reviewed the movie at a specific time after movie release. However, we devise an indirect approach to test this confounding effect by leveraging the different production years of movies in the dataset. It is more likely for reviewers to have watched and rated a movie after it has received Academy Awards if the film was produced many years in the past. For example, a movie like "Toy Story", released in 1995, provided users with nearly thirty years to rate it, which is not the case for more recent movies, such as "Joker", released in 2019. In Appendix Table B5, we present the specifications in Columns (3) and (6) of Table 2 for movies produced in three different decades spanning our dataset: 1995-2000, 2000-2010, and 2010-2019. The results consistently show that the effect remains constant over these three decades, suggesting that the confounding effect related to when users watch movies (before or after the award dates) does not seem to affect our results much.

5.1.2 Other Signals of Quality

Experienced users might be more likely to consider various other aspects related to movie features, such as directors or the main cast. To explore this hypothesis, we adopt an alternative proxy for movie quality: The number of Academy awards and nominations received by the director or main star in the years preceding the movie's production.

In Appendix Table B6, we apply the same specifications used in Table 2 but incorporate this new proxy for movie quality. The results indicate that Top1000 users are more likely to rate movies featuring a nominated or awarded director or main star, reinforcing the idea that Experienced users are indeed better at

recognizing various signals of movie quality.

In line with the previous analysis, we conduct additional checks to ensure the robustness of our findings. Appendix Tables B7, B8, and B9 demonstrate that the results remain consistent: 1) when we use a categorical variable instead of the count of nominations and awards. Specifically, instead of considering the actual number of Oscar nominations and awards, we differentiate between movies based on whether the director or main star has been nominated or awarded for one or more Oscars; 2) when we focus on movies with no j-shaped and u-shaped rating distributions; and 3) when we use the proportions of ratings from the top 1% of MovieLens users or from users over 45 on IMDb.

Restricting our attention to Academy nominations and awards, our measures of quality have been quite limited in scope. As Table 1 indicates, only a minority of movies are either nominated or awarded by the Academy, or are directed or feature stars who have received Academy recognition in the past. However, the vast majority of movies fall outside these categories. Assuming there exists a quality difference among movies that have not received Academy nominations, we aim to validate our earlier findings and determine whether high-quality films within this subset are more likely to be rated by experienced users.

To do this, we introduce two additional measures of quality that allow us to distinguish among movies in the subset which constituted the group of non-high-quality movies in the previous regressions. Firstly, consider a wide range of movie festivals and industry awards, and collect data for winning and nominated movies in our sample. Secondly, we collect each movie's Metascore, which indicates whether the movie was well received by professional critics.

In Appendix Table B10, we implement the specifications presented in Columns (3) and (6) from Table 2, incorporating these two different quality proxies. We focus on the sample of movies that did not receive any Academy Awards (Columns (1) and (4)), excluding those with previously Academy-nominated or awarded directors (Columns (2) and (5), or main stars (Columns (3) and (6)). Once again, the results confirm that a higher proportion of Top1000 users tend to rate higher quality movies. Importantly, this finding holds true even when we broaden our analysis beyond those films considered of exceptionally high quality (i.e., those awarded by the Academy, or with awarded directors and stars).

5.2 Rating Heterogeneity

Next, we show that Experienced users tend to be more stringent in their ratings compared to Novices. In Table 1 we already presented some findings that support this observation: the IMDb average rating of the

Top1000 users is lower than the overall IMDb average rating. To quantify the difference in stringency between Experienced and Novice users, we conduct a regression analysis comparing the average ratings posted by different user groups for the same movies. The estimating equation employed is as follows:

$$\bar{r}_{ij} = \theta_i + \beta_1 Top1000_j + \beta_2 \mathbf{X}_i + \varepsilon_{ij}, \tag{12}$$

where $\bar{r}_{ij}$ represents the average rating for movie i and user group $j = \{\text{Top1000}; \text{Non-Top1000}\}$; θ_i denotes the movie fixed effect; $Top1000_j$ is a binary variable that equals 1 if the user group is the Top1000 IMDb users and 0 otherwise; $\mathbf{X}_i$ encompasses a set of other control variables at the movie level (which can only be identified without the movie fixed effect); and ε_{ij} represents the error term. Table 3 displays four distinct specifications of Equation 12 in which we vary the number of controls. In Columns (1), (2), and (3), we use different sets of controls for movie characteristics, while Column (4) includes movie fixed effects. Across all specifications, we observe a statistically significant difference in average ratings between user groups: Top1000 users post ratings that are consistently over half a star lower compared to Non-Top1000 users. This difference in ratings is not only statistically significant but also economically meaningful, as it accounts for nearly 10% of the average IMDb movie ratings and half of a standard deviation of the IMDb rating distribution.

The higher stringency of the Top1000 users remains negative and statistically significant even when restricting the analysis to movies with no j-shaped rating distributions and different movie genres (see Appendix Tables B11 and B12, respectively). These findings suggest that the observed result is not driven by a specific group of movies. Furthermore, the influence of experience on users' stringency extends to MovieLens as well. In Appendix Table B13, we replicate the analysis using two groups: the top 1% of MovieLens users with the highest number of posted ratings and the bottom 99%. In line with the previous analysis on IMDb, the difference in stringency between Experienced and Novice users is statistically significant and it accounts for approximately 10% of the average rating posted on MovieLens. As a complementary approach, we also employ age as a proxy for experience and investigate the differences in average ratings between users over 45 and under 45 on IMDb. The results are presented in Appendix Table B14 and once again demonstrate that more experienced users – in this case, more senior users – tend to assign lower ratings. To reinforce this idea, Figure 2 illustrates the ratings of both Top1000 and Non-Top1000 users for the 9,426 movies in our sample. The figure reveals that Top1000 users consistently give lower ratings for

Table 3. IMDb Top1000 Users Are More Stringent in Their Ratings

	(1)	(2)	(3)	(4)
$Top1000_j$	-0.570***	-0.570***	-0.570***	-0.570***
	(0.011)	(0.011)	(0.011)	(0.004)
Movie Runtime	0.222***	0.221***	0.199***	
	(0.013)	(0.013)	(0.013)	
$n_i^{critics}(\times 1000)$	1.537***	1.327***	1.286***	
	(0.055)	(0.063)	(0.061)	
$n_i^{Movielens}(\times 1000)$	0.050***	0.054***	0.053***	
	(0.001)	(0.002)	(0.001)	
p_i^{female}			-0.761***	
			(0.054)	
p_i^{US}			-1.066***	
			(0.047)	
Constant	5.888***	5.906***	6.428***	6.521***
	(0.023)	(0.023)	(0.029)	(0.003)
Genre FE	✓	✓	✓	
Year FE		✓	✓	
Movie FE				✓
R^2	0.345	0.348	0.374	0.949
N	18,852	18,852	18,852	18,852

Note: The outcome variable is the IMDb Top1000 and Non-Top1000 average ratings. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

approximately 98% of the movies, despite the ratings showing a remarkably high correlation of 0.89.[33] We also extend this analysis to MovieLens using the top 1% and Bottom 99% of users based on the number of posted ratings in Appendix Figure B4a. The results are consistent, with the top 1% users assigning lower ratings for the majority of movies. Similarly, when considering age as a proxy for experience, we find comparable results. Appendix Figure B4b, for IMDb users over 45 and under 45, exhibits a similar pattern, with experienced users consistently assigning lower ratings for most movies.

5.2.1 Stringency and Diagnosticity

Up to this point, we have measured the different levels of stringency between experienced and novice users. In the terminology of the model described in Section 2, Experienced users' ratings are expressed as $\mathscr{R}_E(Q) = Q + \varepsilon_r$, while Novice users' ratings are $\mathscr{R}_N(Q) = \alpha Q + \beta + \varepsilon_r$. As a result, we have confirmed the significance of the positive coefficient β.

[33] To address potential biases arising from selection into reviewing, in Appendix Figure B3 we conduct a similar analysis for movies with no j-shaped rating distribution. Focusing on this restricted set of movies, Top1000 users continue to post lower ratings for over 98% of the movies.

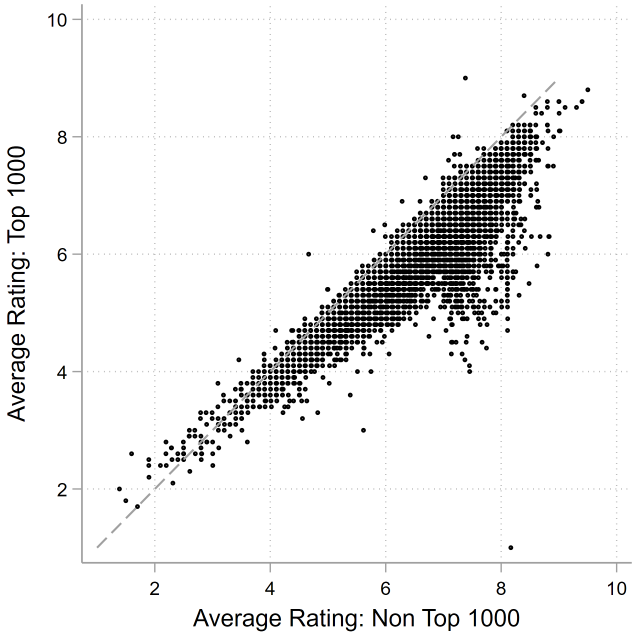


Figure 2. Average Ratings for Top1000 and Non-Top1000 IMDb Users for all movies.

However, experience may also impact the ability to discern different quality levels, penalizing low-quality products and rewarding high-quality ones. To investigate this possibility, we conduct a regression analysis similar to the previous one. This time, we allow the coefficient β_1 in Equation 12 to vary for movies with different quality, proxied with the number of nominations and awards received by the movie, director, and main star. The results of four different interactive effects are presented in Table 4.

Overall, our findings suggest that Experienced users have slightly better diagnostic abilities than Novices. However, the magnitude of the interactive term is an order of magnitude smaller than the average difference in ratings between Top1000 and non-Top1000 users.[34] In fact, the differences in stringency that we find are so large that Top1000 users give lower average ratings to movies with Academy nominations and awards, as well as to movies with awarded directors and stars, compared to Non-Top1000 users' ratings for non-awarded movies. In Appendix Table B15, we replicate the same analysis using MovieLens data, examining the rating difference between the top 1% and bottom 99% of users based on the number of posted ratings. The results show a positive interactive effect of quality, similar to our previous findings. Again, this effect is of an order of magnitude smaller compared to the difference in average ratings between the top 1% and

[34]Given this, the widespread platform policy of overweighting experienced users' ratings compared to those of novices – discussed theoretically in Corollary 1 – is likely to backfire in this setting: the cons ("apples-with-oranges" aggregation resulting from different stringencies) outweigh the pros (higher weight to slightly more diagnostic reviews).

Table 4. IMDb Top1000 Users Are More Stringent in Their Ratings (Different Movie Quality)

	(1)	(2)	(3)	(4)	(5)
$Top1000_j$	-0.570***	-0.578***	-0.573***	-0.585***	-0.579***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
$Top1000_j \times$ Academy Nominations		0.048***			
		(0.007)			
$Top1000_j \times$ Academy Awards			0.047***		
			(0.011)		
$Top1000_j \times$ Main Star Academy Nom.				0.029***	
				(0.003)	
$Top1000_j \times$ Director Academy Nom.					0.020***
					(0.003)
Constant	6.521***	6.521***	6.521***	6.521***	6.521***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Movie FE	✓	✓	✓	✓	✓
R^2	0.949	0.949	0.949	0.950	0.949
N	18,852	18,852	18,852	18,852	18,852

Note: The outcome variable is the IMDb Top1000 and Non-Top1000 average ratings. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

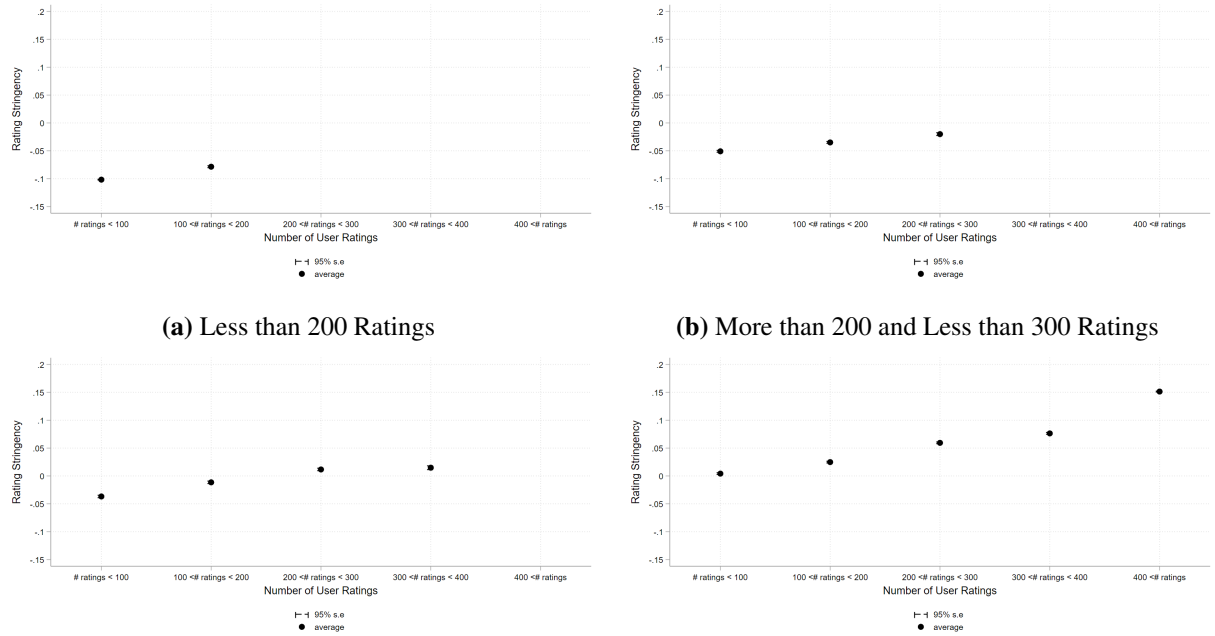
bottom 99%.

Overall, we believe our findings offer striking evidence that the magnitude (and significance) of β is much larger than the deviation of α from 1. Therefore, the disparity between Experienced and Novice users' ratings is primarily driven by their different levels of stringency, rather than by differences in their ability to diagnose quality differences. Given our Propositions 1 and 2, this (combined with choice heterogeneity) suggests the widespread presence of rating compressions, as well as the possibility for ranking reversals. This is exactly what we find, using a different and complementary approach, in Section 6.

5.3 Building Stringency Over Time

In both the previous Section (and in our model), experience has been treated as a fixed, discrete characteristic. This limitation primarily stems from the IMDb dataset, which only offers a dichotomy between Top1000 and non Top1000 users, and does not allow us to track individual users' behavior over time. In reality, users are likely to develop experience gradually.

To validate this theory, and thus strengthen our intuition that experience does, indeed, lead to higher stringency, we exploit the entire MovieLens dataset, which contains 25 million ratings from 32,202 users. By analyzing users' rating histories in this dataset, we can track the evolution of users' rating stringency



(a) Less than 200 Ratings

(b) More than 200 and Less than 300 Ratings

(c) More than 300 and Less than 400 Ratings

(d) More than 400 Ratings

Figure 3. MovieLens User Stringency over the Number of Ratings

over time. To measure the stringency of a particular rating given by user j for a movie i, we calculate the difference between the average rating given by all users for that movie $\bar{r}_i$ and the user's rating r_{ij}: $s_{ij} = \bar{r}_i - r_{ij}$. Positive values of s_{ij} indicate lower-than-average ratings, implying that the user who posted the rating is more stringent than others who rated the same movie.[35] Next, we segment the user sample into four groups based on the total number of ratings they have posted on the platform: users with less than 200 ratings, users with 200 to 300 ratings, users with 300 to 400 ratings, and users with more than 400 ratings.

Figure 3 illustrates the average rating stringency plotted against the number of ratings posted for each of these four user groups. We observe a clear trend of rating stringency increasing with the accumulation of more ratings across all four user groups: users become more stringent as they gain experience by watching and rating more movies. Interestingly, users who post more reviews tend to have higher stringency levels in all of their reviews – including their very first ones.

There are two potential reasons for this, one related to nurture, the other to nature. First, users are likely to have prior movie-watching experience before joining MovieLens, which may explain the initial differences in stringency levels as a result of accumulated experience prior to joining the platform (if the

[35] In the next Section, we refine this approach further and describe an algorithm to mitigate biases in aggregate ratings and improve the quality of signals provided.

latter correlates with the level of activity once joining the platform, which seems plausible). This, combined with the downward trend in leniency in individual MovieLens ratings described above, suggests a nurture explanation. Second, some users may have an inherently stricter internal scale for a given level of experience (nature).

6 De-biasing Algorithm

6.1 Description

In this Section, we develop a simple recursive algorithm to mitigate the biases in ratings we have described.

Our algorithm exploits the full history of users' ratings to quantify each user's stringency level and to normalize movies' ratings accordingly. More formally, for each movie i, we compute the movie-user specific stringency as the average movie rating, $\bar{r}_i$, minus the rating posted by the user, r_{ij}. In Section 5.3 we showed that users' stringency tends to increase as they post more ratings on the platform. In line with these findings, we do not calculate a single average stringency level for each user. Instead, we use the previously established thresholds to calculate average stringency levels for each user across different rating groups. These groups include the first 100 ratings, ratings between the 100th and 200th, ratings between the 200th and 300th, ratings between the 300th and 400th, and all subsequent ratings beyond the 400th.

By grouping a user's ratings into these categories, we effectively treat them as if they were posted by four different users, each with a different stringency level. Accordingly, for each group of ratings k, we define the set of movies in the group as $\mathscr{I}_k$. Then the user's stringency for that group of ratings is the average of the movie-user specific stringency over the set $\mathscr{I}_k$:

$$s_k := \frac{\sum_{i \in \mathscr{I}_k} \left(\bar{r}_i - r_{ik}\right)}{n_k}, \tag{13}$$

where n_k is the cardinality of $\mathscr{I}_k$.

Then, we compute new movie ratings which take into account – and correct for – how stringent users are. We define by $\mathscr{K}_i$ the set of all users who have watched movie i, and by n_i its cardinality. Then, we update, or normalize, movie i's rating $\bar{r}_i$ to:

$$\bar{r}_i^{debias} := \frac{\sum_{k \in \mathscr{K}_i} \left(r_{ik} + s_k\right)}{n_i} = \bar{r}_i + \frac{\sum_{k \in \mathscr{K}_i} s_k}{n_i}. \tag{14}$$

This correction is, in many ways, "mechanical". We see this as an appealing feature. In particular, it weights all opinions equally, and similarly, it does not require us to make assumptions about the rating and choice processes for each category of users.

Equations 13 and 14 are clearly interdependent. Each individual user's stringency depends on how our normalization affects the ratings of the movies she has watched, and this normalization in turn depends on the average individual stringency among all users watching the movie. To deal with this interdependence, we solve this system of equations recursively, by iterating the process forward until convergence to its fixed point.

With this algorithm, we aim to remove reviewer heterogeneity from ratings. Thus, our approach shares the same intent with using products and reviewer fixed effects in line with the analysis by Dai, Jin, Lee and Luca (2018). The main difference between these two approaches lies in the information about the network connecting movies and moviegoers, and the fixed-point equilibrium analysis it allows us to perform. In particular, using product and reviewer fixed effects cannot fully account for the fact that an experienced reviewer could mostly rate movies that have been rated by other users with a similar degree of experience. In this scenario, a movie reviewer fixed effect will assign a relatively low stringency level to this user. However, given that the network in our data is a rich one, we can specifically address this problem with our algorithm. Here, each user's stringency is not entirely dependent on the reviewers who watch the *same* movies. In our algorithm, each step incorporates more information about the entire network, as we update the stringencies of users and movies over time until convergence.

6.2 Results

How do our de-biased ratings compare to the original ones? Which movies were more penalized / aided by the aggregation bias we describe? We answer this question, we examine the difference between the our de-biased average movie ratings and the original MovieLens ones, $\bar{r}_i^{debias} - \bar{r}_i$, for the sample of 9,426 movies used in the previous Sections. We regress this difference on proxies for movie quality, user experience, and a set of control variables in line with Equation 11. Our results are reported in Tables 5 and 6. In both Tables, we use the same specifications shown in Tables 2, B6, and B10.

In Table 5, we use the number of Academy nominations and awards as proxies for quality. We find that nominated and awarded movies are penalized by the rating system: the difference $\bar{r}_i^{debias} - \bar{r}_i$ is larger for movies that received Academy Awards (despite their $\bar{r}_i$'s being larger, on average, in the first place). This

Table 5. Winners and Losers of Our De-Biasing Algorithm: High-Quality Movies

	(1)	(2)	(3)	(4)	(5)	(6)
Academy Nominations	0.002	0.002*	0.002**			
	(0.001)	(0.001)	(0.001)			
Academy Awards				0.006***	0.005***	0.005***
				(0.002)	(0.002)	(0.002)
Movie Runtime	-0.014***	-0.013***	-0.012***	-0.014***	-0.013***	-0.012***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
$n_i^{critics}(\times 1000)$	-0.096***	-0.153***	-0.162***	-0.096***	-0.152***	-0.161***
	(0.007)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
$n_i^{Movielens}(\times 1000)$	-0.004***	-0.003***	-0.003***	-0.005***	-0.003***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
p_i^{female}			-0.163***			-0.163***
			(0.006)			(0.006)
p_i^{US}			-0.007			-0.007
			(0.005)			(0.005)
Constant	1.109***	1.113***	1.147***	1.109***	1.113***	1.147***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.187	0.232	0.282	0.188	0.232	0.283
N	9,426	9,426	9,426	9,426	9,426	9,426

Note: The outcome variable is the difference between the de-biased movie ratings and the average movie ratings on MovieLens ($\bar{r}_i^{debias} - \bar{r}_i$). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

result is less pronounced for movies that received Academy nominations.

Moreover, thanks to the nature of the MovieLens dataset, we can test whether this penalization is due to experienced users. This is equivalent to testing whether, in light of our de-biasing, experience correlates with stringency, which would lend additional confirmation to our findings in Section 5.2. In Table 6 we measure the relationship between the gain (or loss) for a movie following our algorithm, $\bar{r}_i^{debias} - \bar{r}_i$, and two measures of user experience: *i*) the average number of ratings that MovieLens users posted before rating movie i ($\bar{n}_i^{users}$), *ii*) the average total number of ratings posted by MovieLens users who rated movie i at some point in time ($\bar{n}_i^{tot\ users}$).[36]

We use these two variables to proxy users' experience as we have shown in Section 5.3 that stringency appears to be positively correlated with the overall number of ratings posted on the platform and also increases over time with the number of ratings posted by each user. The results in Table 6 confirm that movies

[36] In other words, if user j's 50th rating is to movie i, and user j has a total of 100 ratings on MovieLens, we have $\bar{n}_{ij}^{users} = 50$ and $\bar{n}_{ij}^{tot\ users} = 100$. We then average over all users who have watched movie i (at any point in time) to obtain $\bar{n}_i^{users}$ and $\bar{n}_i^{tot\ users}$.

Table 6. Winners and Losers of Our De-Biasing Algorithm: Movies Rated by Experienced Users

	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{n}_i^{users}(\times 1000)$	0.148***	0.186***	0.182***			
	(0.003)	(0.003)	(0.003)			
$\bar{n}_i^{tot\ users}(\times 1000)$				0.119***	0.128***	0.123***
				(0.002)	(0.002)	(0.002)
Movie Runtime	-0.006***	-0.004***	-0.005***	-0.004***	-0.003***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$n_i^{critics}(\times 1000)$	-0.090***	-0.017**	-0.026***	-0.052***	-0.017***	-0.027***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
$n_i^{Movielens}(\times 1000)$	-0.000***	-0.001***	-0.001***	-0.000**	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
p_i^{female}			-0.108***			-0.097***
			(0.006)			(0.006)
p_i^{US}			-0.051***			-0.041***
			(0.005)			(0.005)
Constant	0.997***	0.962***	1.005***	0.960***	0.947***	0.986***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.400	0.442	0.473	0.434	0.453	0.476
N	9,426	9,426	9,426	9,426	9,426	9,426

Note: The outcome variable is the difference between the de-biased movie ratings and the average movie ratings on MovieLens ($\bar{r}_i^{debias} - \bar{r}_i$). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

rated by users who have posted more ratings (before rating the movie and overall) are significantly penalized by the current rating system. As a robustness check, Appendix Table B16 presents the same analysis using as a proxy for users' "experience" the ratio between the number of ratings posted by the Top1000 IMDb users and Non-Top1000 users ($\pi_i^{Top1000}$). The results confirm that movies watched by a larger proportion of IMDb Top1000 users are more penalized by the rating aggregation on MovieLens.

It is crucial to emphasize that our algorithm is fully agnostic about both the choice and rating process. Specifically, it does not assume a positive correlation between experience and stringency; nor does it assume that experienced users choose better movies. Thus, the previous results serve as an additional validation of our rating heterogenity and choice heterogenity results presented in Sections 5.1 and 5.2.

6.3 Measuring Ranking Reversals

By giving us a proxy for quality (de-biased ratings), our algorithm also allows to quantify the presence of rating compressions and ranking reversals. To do so, for every pair of movies i and l, we calculate the

product of the difference between their original average ratings $(\bar{r}_i - \bar{r}_l)$ and the difference between their new de-biased ratings $(\bar{r}_i^{debias} - \bar{r}_l^{debias})$. This allows us to determine whether a pair of movies is affected by ranking reversals. Specifically, when the product $(\bar{r}_i - \bar{r}_l)(\bar{r}_i^{debias} - \bar{r}_l^{debias})$ is less than zero, it indicates a situation where the original ratings would have suggested that one movie is of higher quality than the other, while our de-biased ratings suggest this might have simply been due to the more stringent set of users rating movie i.

We calculate this product for all the $87,143,550$ potential combinations among the $9,426$ movies included in our study. Approximately 8.4% of these combinations exhibit rankings reversals. This approach also includes cases with very small differences in average ratings as ranking reversals. For example, if the difference between the original average ratings of two movies, A and B, was $4.45 - 4.44 = 0.01$, and the difference between the de-biased average ratings was $4.44 - 4.45 = -0.01$, then the product $(\bar{r}_A - \bar{r}_B)(\bar{r}_A^{debias} - \bar{r}_B^{debias})$ would be equal to -0.0001, which we classify as a case of ranking reversal. Nevertheless, due to potential approximations or measurement errors, such a tiny difference should not be considered a significant case of ranking reversal.

To address this concern, we also demonstrate that a substantial number of ranking reversals exist in our dataset even when we only consider cases in which the product $(\bar{r}_i - \bar{r}_l)(\bar{r}_i^{debias} - \bar{r}_l^{debias})$ is less than -0.01. Here, we still observe 2.5% ranking reversals in all combinations of our movies. While -0.01 might seem small, it is important to note that a threshold of -0.01 implies a minimum value of $(\bar{r}_i - \bar{r}_l) - (\bar{r}_i^{debias} - \bar{r}_l^{debias})$ equal to 0.2,[37] which is equivalent to 40% of the MovieLens ratings standard deviation. Similar trends emerge when we focus on a subset of movies categorized within specific genres. In particular, 7%, 7.8%, and 10.6% of the rank reversals occur within combinations of movies that fall into the action, comedy, and drama genres, respectively.

These percentages provide a broad perspective on the prevalence of ranking reversals within our dataset. However, we can delve deeper into our analysis to examine how different movies are affected by ranking reversals. Specifically, for each movie, we can compute the proportion of combinations with the other 9,425 movies that are affected by ranking reversals (i.e., where $(\bar{r}_i - \bar{r}_l)(\bar{r}_i^{debias} - \bar{r}_l^{debias}) < 0$). In doing so, we simultaneously measure the extent to which a movie is disadvantaged or advantaged by the original ratings relative to all other movies.

[37] Given $x > 0$ and $y < 0$ such that $xy = -0.01$, the difference $x - y$ is minimized when $x = 0.1$ and $y = -0.1$, and thus at $x - y = 0.2$. Conversely, one could have a case where original ratings are very close ($y = -0.01$) while the de-biased ratings are not ($x = 1$), which results in the same product of -0.01 but a much larger relative bias of $x - y = 1.01$.

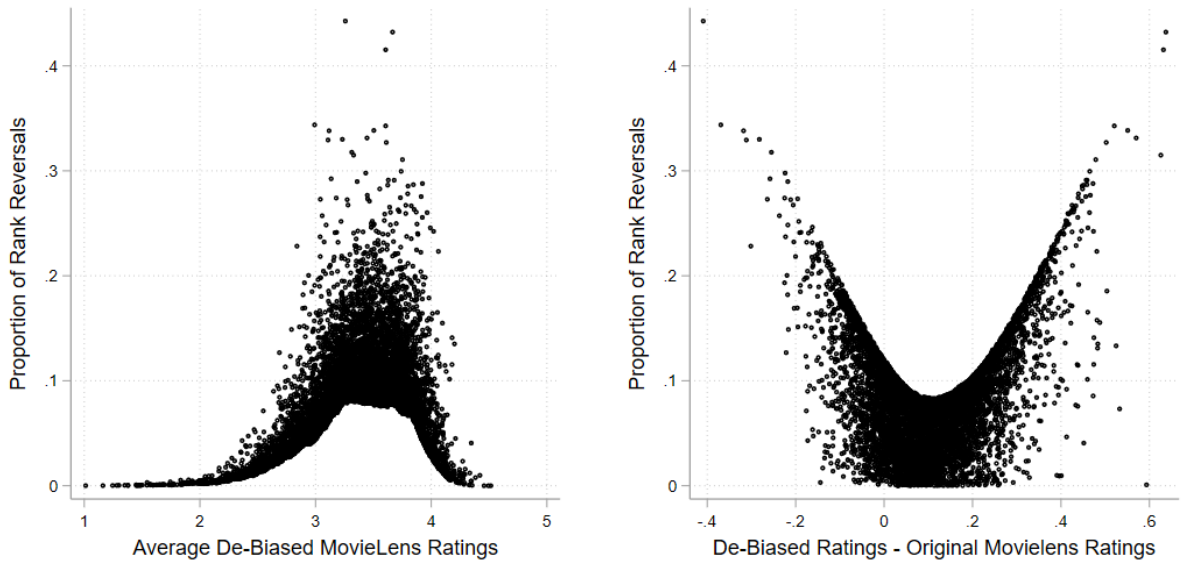


Figure 4. Proportion of Ranking Reversals for Movies with Different De-biased Ratings and Different Premia from De-biasing Algorithm

Figure 4 shows a scatterplot illustrating the variation in rank reversals among movies with different de-biased ratings ($\bar{r}_i^{debias}$) and different gaps between the de-biased ratings and their original counterparts ($(\bar{r}_i - \bar{r}_i^{debias})$). Movies with extremely low or high de-biased ratings show minimal susceptibility to ranking reversals, while those with moderate to high de-biased ratings (between 3 and 4 stars out of 5) are more susceptible to reversals. This finding underscores the importance of ranking reversals in users' decision-making processes. In particular, movies of very low or high quality are relatively easy for users to recognize as either "must avoid" or "must watch". For movies of average quality, however, rankings serve as crucial signals that can influence users' decisions. The observed frequency of ranking reversals for these average-quality movies suggests that the rating biases presented in this paper are particularly relevant in cases where ratings should be more informative ad they are more consequential.

Moreover, the highest proportion of ranking reversals occurs for movies with large values of $(\bar{r}_i - \bar{r}_i^{debias})$, be it negative or positive: a movie with ranking reversals is either one that has been unfairly downgraded by the original ratings or excessively praised because it attracted less stringent users.

These patterns remain consistent when we conduct the same analyses using a -0.01 threshold for identifying ranking reversals (Appendix Figure B5). In this scenario, movies with de-biased ratings between 3 and 4, as well as movies with large values of $(\bar{r}_i - \bar{r}_i^{debias})$, still exhibit the highest incidence of ranking reversals.

It is interesting to see which movies, ex-post, are most rewarded (punished) by our de-biasing procedure. Among the biggest winners of our correction we find, for example, "Destino", an animated surrealist short film that originated from a collaboration between Walt Disney and Salvador Dalí which was nominated for the 2003 "Best Animated Short Film" Academy Award and "The Unknown Girl," a French mystery film that was selected for the Palme d'Or at the 2016 Cannes Film Festival. Among the biggest losers, we find "Zenon: Z3", a Disney Channel television movie that was panned on Rotten Tomatoes (and is altogether missing on Metacritic), and "The Choice", a movie with a relatively wide distribution and good box office performance, but which received negative reviews from critics (a Metacritic score of 23 out of 100).[38]

7 Conclusion

We investigate the consequences of consumer heterogeneity on online consumer ratings. We argue that better products are systematically purchased by a greater share of more experienced consumers, who on average post more stringent evaluations. As a result, ratings compress, and sometimes reverse, products' qualities. We test our claims by using data from IMDb and MovieLens and find striking support for them. Experienced users rate higher quality movies, as proxied by both nominations and awards to international festival and industry awards and movie critics reviews. Moreover, they rate movies much more stringently. This is not just true on average: the relationship holds for a remarkable 98% of movies in our sample, irrespective of genre, year of production, and popularity.

Combined, these two facts imply a bias against higher quality movies. This bias is highly problematic, as the primary role of ratings is arguably that of enabling consumers to separate the best products from their inferior alternatives. This result is particularly surprising given that some of the high quality movies we identify in the empirical analysis are those that perform well in global film festivals. These movies are sometimes considered a genre in their own right, i.e. "arthouse" or "foreign", and thus might attract a highly self-selected set of consumers who would plausibly give them more positive ratings. The fact that the opposite happens adds extra weight to our argument that these consumers simply rate on a stricter scale.

Next, we show that, however pervasive and problematic this bias is, correcting for it is rather straightforward. We de-bias the ratings by exploiting the full history of users' ratings and mechanically equating users' stringency levels. Our correction does not require us to overweight some ratings and discard others,

[38] For statistics about "The Choice" see: https://www.the-numbers.com/movie/Choice-The-(2016)#tab=summary.

or to have a prior of which movies are actually of high quality. This approach leads to normalized aggregate ratings that better correlate with external proxies of quality.

In thinking about the generalizability of our results, it is worth pointing out that, if anything, the movie market is fairly "egalitarian", partly by virtue of its uniform prices: certain – arguably high quality – movies (*e.g. "The Shawshank Redemption"*, *"Oppenheimer"* or *"Schindler's List"*) are watched by a majority of consumers of very different experience levels. We believe that the bias we identify will be even stronger when looking at product categories (be it digital cameras, restaurants, or hotels) in which the discrepancy in choices, and we suspect ratings, between consumers is much more pronounced.[39] With restaurants, and many other categories, we suspect aggregate ratings likely reflect an "apple-with-oranges" comparison even more so than they do with movies.

Industry players seem to at least partly understand these dynamics. Some ratings platforms are starting to internalize the idea that their users are very heterogeneous in their stringency, and trying to correct for it. Similar to our approach, BeerAdvocate, a popular beer ratings platform, attributes a stringency score called *rDev* to each of its users, based on their reviews.[40] However, unlike our proposed algorithm, BeerAdvocate stops short of computing (and correcting for) product-specific stringency scores, which makes internalizing this information extremely difficult for consumers.

We believe our results are important for managers, platforms and consumers alike. For managers, we suggest the possibility of segmentation based on which consumers are most likely to generate positive word of mouth, boosting future demand. For example, a movie that receives early ratings by the most experienced IMDb users is much less likely to "look good" than one reviewed by casual movie watchers. Instead of targeting the most experienced consumers, managers should try and obtain more lenient ratings to kick-start their products' success.

For platforms, we believe our results warrant caution with overemphasizing the opinions of "super reviewers", an increasingly widespread practice (e.g., Amazon, IMDb, Yelp).[41] As this elite group of reviewers is small and rates on a different (and harsher) scale from everyone else, emphasizing their ratings without normalizing their scale likely exacerbates the "apples-with-oranges" bias we highlight.

[39] The main empirical disadvantage presented by restaurant ratings is the key role played by prices: ratings need not solely represent quality. Thus, the fact that, say, Michelin star restaurants are held to a higher standards than their non-Michelin counterparts would not necessarily be evidence for the mechanism we propose: even the same consumer's ratings could reflect differences in $Q - P$ and thus compress differences in Q. See Luca and Reshef (2021).

[40] See https://www.beeradvocate.com/community/threads/beeradvocate-ratings-explained.184726/.

[41] For example, see the Yelp Elite Squad initiative: https://www.yelp.ca/elite.

For consumers, we highlight an important source of mislearning from reviews, and suggest that thinking about the self-selection of raters for each product is key to unbiased learning (*e.g., "Is this movie / restaurant / hotel / book likely to attract a crowd of experienced, strict reviewers, which makes it look worse than it actually is? Or is it the other way around?"*). However, drawing such inferences is complicated by the fact that it relies on consumers having some prior knowledge about the product, which they may lack, especially if they are seeking ratings to inform their decision-making. At the very least, when individual reviews are accessible (as is typically the case on most online platforms), consumers should consider that a lukewarm or negative review from a "superstar" reviewer with thousands of ratings on the platform may not necessarily indicate a significant issue (and conversely, an enthusiastic review from a novice might not guarantee exceptional quality).

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar (2022) 'Learning from reviews: The selection effect and the speed of learning.' *Econometrica* 90(6), 2857–2899

Akesson, Jesper, Robert W Hahn, Robert D Metcalfe, and Manuel Monti-Nussbaum (2023) 'The impact of fake reviews on demand and welfare.' Technical Report, National Bureau of Economic Research

Alba, Joseph W, and J Wesley Hutchinson (1987) 'Dimensions of consumer expertise.' *Journal of consumer research* 13(4), 411–454

Anderson, Eric T, and Duncan I Simester (2014) 'Reviews without a purchase: Low ratings, loyal customers, and deception.' *Journal of Marketing Research* 51(3), 249–269

Anderson, Michael, and Jeremy Magruder (2012) 'Learning from the crowd: Regression discontinuity estimates of the effects of an online review database.' *The Economic Journal* 122(563), 957–989

Aridor, Guy, Duarte Gonçalves, Daniel Kluver, Ruoyan Kong, and Joseph Konstan (2022) 'The economics of recommender systems: Evidence from a field experiment on movielens.' *arXiv preprint arXiv:2211.14219*

Arrow, Kenneth J (1962) 'The economic implications of learning by doing.' *The review of economic studies* 29(3), 155–173

Besbes, Omar, and Marco Scarsini (2018) 'On information distortions in online ratings.' *Operations Research* 66(3), 597–610

Blyth, Colin R (1972) 'On simpson's paradox and the sure-thing principle.' *Journal of the American Statistical Association* 67(338), 364–366

Bondi, Tommaso (2023) 'Alone, together: A model of social (mis)learning from consumer reviews'

Brandes, Leif, David Godes, and Dina Mayzlin (2022) 'Extremity bias in online reviews: The role of attrition.' *Journal of Marketing Research* 59(4), 675–695

Cabral, Luis (2012) 'Reputation on the internet.' *The Oxford Handbook of the Digital Economy* pp. 343–354

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

Chevalier, Judy, Yaniv Dover, and Dina Mayzlin (2014) 'Promotional reviews: An empirical investigation of online review manipulation.' *American Economic Review* 104(8), 2421–55

Crawford, Vincent P, and Joel Sobel (1982) 'Strategic information transmission.' *Econometrica* pp. 1431–1451

Dai, Weijia, Ginger Jin, Jungmin Lee, and Michael Luca (2018) 'Aggregation of consumer ratings: an application to yelp. com.' *Quantitative Marketing and Economics* 16, 289–339

De Langhe, Bart, Philip M Fernbach, and Donald R Lichtenstein (2016) 'Navigating by the stars: Investigating the actual and perceived validity of online user ratings.' *Journal of Consumer Research* 42(6), 817–833

Dellarocas, Chrysanthos, and Ritu Narayan (2006) 'A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth.' *Statistical Science* 21(2), 277–285

Ellison, Glenn, and Drew Fudenberg (1995) 'Word-of-mouth communication and social learning.' *The Quarterly Journal of Economics* 110(1), 93–125

Ericsson, K Anders, Ralf T Krampe, and Clemens Tesch-Römer (1993) 'The role of deliberate practice in the acquisition of expert performance.' *Psychological review* 100(3), 363

Esponda, Ignacio (2008) 'Behavioral equilibrium in economies with adverse selection.' *The American Economic Review* 98(4), 1269–1291

Fainmesser, Itay P, Dominique Olié Lauga, and Elie Ofek (2021) 'Ratings, reviews, and the marketing of new products.' *Management Science* 67(11), 7023–7045

Fang, Limin (2022) 'The effects of online review platforms on restaurant revenue, consumer learning, and welfare.' *Management Science* 68(11), 8116–8143

Farronato, Chiara, and Georgios Zervas (2022) 'Consumer reviews and regulation: evidence from nyc restaurants.' Technical Report, National Bureau of Economic Research

Gerardi, Dino, and Leeat Yariv (2008) 'Costly expertise.' *American Economic Review* 98(2), 187–193

Godes, David, and José C Silva (2012) 'Sequential and temporal dynamics of online opinion.' *Marketing Science* 31(3), 448–473

Greenleaf, Eric A (1992) 'Improving rating scale measures by detecting and correcting bias components in some response styles.' *Journal of Marketing Research* 29(2), 176–188

Harper, F Maxwell, and Joseph A Konstan (2015) 'The movielens datasets: History and context.' *ACM Transactions on Interactive Intelligent Systems (TIIS)* 5(4), 1–19

He, Sherry, Brett Hollenbeck, and Davide Proserpio (2022) 'The market for fake reviews.' *Marketing Science* 41(5), 896–921

Hu, Nan, Paul A Pavlou, and Jie Jennifer Zhang (2009) 'Why do online product reviews have a j-shaped distribution? overcoming biases in online word-of-mouth communication.' *Communications of the ACM* 52(10), 144–147

Hu, Nan, Paul A Pavlou, and Jie Zhang (2017) 'On self-selection biases in online product reviews.' *MIS quarterly* 41(2), 449–475

Jacobsen, Grant D (2015) 'Consumers, experts, and online product evaluations: Evidence from the brewing industry.' *Journal of Public Economics* 126, 114–123

Johnson, Kathy E, and Carolyn B Mervis (1997) 'Effects of varying levels of expertise on the basic level of categorization.' *Journal of experimental psychology: General* 126(3), 248

Karaman, Hülya (2021) 'Online review solicitations reduce extremity bias in online review distributions and increase their representativeness.' *Management Science* 67(7), 4420–4445

Kőszegi, Botond, and Matthew Rabin (2006) 'A model of reference-dependent preferences.' *The Quarterly Journal of Economics* 121(4), 1133–1165

Krishna, Vijay, and John Morgan (2001) 'A model of expertise.' *The Quarterly Journal of Economics* 116(2), 747–775

LaTour, Kathryn A, and John A Deighton (2019) 'Learning to become a taste expert.' *Journal of Consumer Research* 46(1), 1–19

Lewis, Gregory, and Georgios Zervas (2016) 'The welfare impact of consumer reviews: A case study of the hotel industry.' *Unpublished manuscript*

Li, Xinxin, and Lorin M Hitt (2008) 'Self-selection and information role of online product reviews.' *Information Systems Research* 19(4), 456–474

Luca, Michael (2016) 'Reviews, reputation, and revenue: The case of yelp. com.' *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*

Luca, Michael, and Georgios Zervas (2016) 'Fake it till you make it: Reputation, competition, and yelp review fraud.' *Management Science* 62(12), 3412–3427

Luca, Michael, and Oren Reshef (2021) 'The effect of price on firm reputation.' *Management Science* 67(7), 4408–4419

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie (2018) 'A theory of non-bayesian social learning.' *Econometrica* 86(2), 445–490

Narasimhan, Chakravarthi, Chuan He, Eric T Anderson, Lyle Brenner, Preyas Desai, Dmitri Kuksov, Paul Messinger, Sridhar Moorthy, Joseph Nunes, Yuval Rottenstreich, Richard Staelin, George Wu, and Z.John Zhang (2005) 'Incorporating behavioral anomalies in strategic models.' *Marketing Letters* 16, 361–373

Nguyen, Peter, Xin Wang, Xi Li, and June Cotte (2021) 'Reviewing experts' restraint from extremes and its impact on service providers.' *Journal of Consumer Research* 47(5), 654–674

Papanastasiou, Yiangos, and Nicos Savva (2017) 'Dynamic pricing in the presence of social learning and strategic consumers.' *Management Science* 63(4), 919–939

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

Rocklage, Matthew D, Derek D Rucker, and Loran F Nordgren (2021) 'Emotionally numb: Expertise dulls consumer experience.' *Journal of Consumer Research* 48(3), 355–373

Rossi, Peter E, Zvi Gilula, and Greg M Allenby (2001) 'Overcoming scale usage heterogeneity: A bayesian hierarchical approach.' *Journal of the American Statistical Association* 96(453), 20–31

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

Spiegler, Ran (2011) *Bounded rationality and industrial organization* (Oxford University Press)

_ (2016) 'Bayesian networks and boundedly rational expectations.' *The Quarterly Journal of Economics* 131(3), 1243–1290

Sun, Monic (2012) 'How does the variance of product ratings matter?' *Management science* 58(4), 696–707

Tadelis, Steven (2016) 'Reputation and feedback systems in online platform markets.' *Annual Review of Economics* 8, 321–340

Winer, Russell S, and Peter S Fader (2016) 'Objective vs. online ratings: Are low correlations unexpected and does it matter? a commentary on de langhe, fernbach, and lichtenstein.' *Journal of Consumer Research* 42(6), 846–849

Yoganarasimhan, Hema (2013) 'The value of reputation in an online freelance marketplace.' *Marketing Science* 32(6), 860–891