The Good, The Bad and The Picky: Reference Dependence and the Reversal of Movie Ratings

Tommaso Bondi, Michelangelo Rossi and Ryan Stevens*
Cornell, Telecom Paris and Meta

July 18, 2022

Abstract

We explore the consequences of referent-dependent preferences on the nature of online reviews. Consumers differ in their experience, which has two effects. First, experience is instrumental to choice: experts purchase better products than non-experts. Second, because of their superior choices, experts endogenously form higher reference points, and post harsher ratings for given quality. Combined, these two facts imply a bias against higher quality products. When this bias gets large, ratings are non-monotonic in quality: higher-quality products can obtain lower ratings than their inferior alternatives. We test our theory using two large datasets obtained from well-known movie rating websites and find strong support for it. We proxy users' expertise with the total number of ratings posted on the platforms. Using external measures of quality, such as the Academy Awards, we show that experts rate movies of higher quality compared to non-experts. Moreover, experts post more stringent ratings for the same movies. Finally, we "debias" ratings exploiting the full history of users' ratings to level up their stringency levels. This approach leads to normalized aggregate ratings that reduce the bias against higher-quality products and are more in line with external measures of quality.

*We thank Andrew Ching, Daniel Csába, Fabrizio Dell'Acqua, Apostolos Filippas, Brett Hollenbeck, Gentry Johnson, Masakazu Ishihara, Michael Luca, Franz Ostrizek, Omid Rafieian, Matthew Rocklage, Lena Song, Daniel Stackman, Shervin Tehrani and particularly Luís Cabral, John Horton and Raluca Ursu for their helpful comments and suggestions. We are deeply grateful to everyone at GroupLens for collecting and sharing the MovieLens data, and especially to Daniel Kluver for helping us navigating it. We also thank seminar participants at the Frank M. Bass FORMS Conference, the London Business School Trans Atlantic Conference, and the NYU Stern Friday Workshop. All errors are our own.

1 Introduction

Online consumer reviews have become a ubiquitous driver of choice. But, to what extent can we trust their informational content? Because consumer reviews largely measure subjective satisfaction, they can reflect characteristics of their writers just as much as of what is being reviewed. If individuals' characteristics correlate with their choices, a self-selection bias in ratings will arise. 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, which has two separate effects. On one hand, more experienced consumers identify and buy higher quality products. On the other hand, the level of satisfaction consumers get from the products they purchase (and, thus, their ratings) depend negatively on their standards, or the level of quality they are used to.

Examples of this phenomenon are ubiquitous: a Michelin restaurant might appear only average to all of its customers who routinely eat at Michelin establishments. As a result, it may obtain worse reviews than a lower quality restaurant attracting less demanding customers. A National Geographic photographer might find its latest, professional-level camera inadequate, and rate it negatively, while many novices will rave about the quality of their first camera, because they compare its photos' quality to that of their smartphones'.

Of course, prices may play a key role in all of the above examples: after all, shouldn't consumers hold more expensive products to a higher standard? To eliminate this concern empirically, we show that this bias severely affects movie ratings, despite the fact that movies are *uniformly priced*.

We first propose a simple theoretical model building on the idea that a combination of heterogenous consumer experience and reference-dependent individual utility (à la Kőszegi and Rabin [2006]) gives rise to a compression of ratings, effectively penalizing high-quality products compared to their lower quality alternatives. This is highly problematic since one of the primary objectives of online reviews systems is to help consumers screen lower quality products.

Furthermore, the model shows that ratings need not only compress quality differences, they can actually reverse quality rankings. We show that whether this occurs depends on the combination of three of the model's primitives: the share of experienced consumers, the importance of standards in shaping utility, and the size of the utility gap due to experience.

To elucidate the phenomenon, consider the following example: movie A is better than movie B (let us assume there is no horizontal differentiation). Consumers are equally split into movie experts and non-experts. 90% of movie experts choose movie A: their experience

is instrumental in identifying the superior option. Non-experts, instead, pick A and B randomly. On average, experts rate A with a score of 7, and B with a score of 6, while Non-Experts rate them 8.5 and 8 respectively. These features resemble our data in close detail: not only do experts review, on average, better movies (where "better" will be carefully formalized in Section 4), but their reviews are on average lower for every movie, and – in line with loss aversion – more so for lower quality levels (in this example, $7-6 > 8.5-8$). Once reviews are aggregated, movie A and B's scores are given by the weighted average of ratings from the two groups, that is, $R_A = \frac{0.9 \cdot 7 + 0.5 \cdot 8.5}{0.9+0.5} = 7.53$ and $R_B = \frac{0.1 \cdot 6 + 0.5 \cdot 8}{0.1+0.5} = 7.66$. That is, both consumer groups enjoy A more than B, but B ends up having a higher score, due to the less stringent standard it faces. Future consumers learning (naïvely) from these ratings would end up choosing B over A, even though they would have enjoyed A more.[1]

We conclude our theoretical analysis by proposing two distinct but equivalent solutions to the problem, and discussing some (real and apparent) remedies observed in the real world. In particular, commonly observed on most ratings platforms – such as *Amazon, eBay* and *Yelp* – are the overweighting of more experienced users' opinions and the inflation of the average ratings of products that received many ratings. We show that while the first backfires, the latter can prove beneficial, provided some assumptions are satisfied.

We empirically substantiate the severity of this bias – as well as its drivers and consequences – by studying consumer movie ratings. In particular, we scraped detailed aggregate data for over 9000 movies from IMDb, a popular movie rating platform, and complemented it with a massive individual-level rating dataset from MovieLens. This is a platform for movie discovery and personalized recommendations created by GroupLens, a research group within the University of Minnesota's Computer Science department.

On top of the advantage given by movies' uniform pricing, the aggregate data we scraped comes in ideal form for our purposes: for each movie, on top of the overall score, an advanced search reveals more detailed averages concerning only specific subgroups of consumers. Of special interest for us is the Top1000 Users category, which groups together the 1000 users who have posted the most ratings on the platform. We will henceforth refer to them as Experts and contrast their behavior with that of all other users (Non-Experts).

Experts watch, on average, higher-quality movies. To proxy movie quality, we employ external sources such as the nominations and awards for the five most relevant Academy awards and consumer ratings on both platforms. Moreover, Experts are harsher than Non-Experts in their ratings. This is not only true on average, but the result holds for over 98% of movies in our sample. On average, Experts leave scores that are almost 6% and 7% lower than Non-Experts for awarded and nominated movies, respectively; and more than 9% lower if we restrict attention to lower-quality, non-nominated movies. The Experts' self-selection based on quality combined with their rating behavior implies that aggregate ratings penalize

[1] Our model can be thought of as an equilibrium version of Simpson's Paradox (Blyth [1972]) applied to online ratings. We thank Nikhil Garg for suggesting this link.

high-quality movies compared to their inferior alternatives.

Last, we "debias" the ratings. The key observation to this end is that however pervasive, this type of self-selection bias allows for a straightforward correction. To do so, we exploit the full history of individual ratings and compute the user stringency. By subtracting this user-specific dimension from each rating, we level up users' stringency. After having normalized Experts and Non-Experts' rating scales, we recompute aggregate ratings using these normalized ratings. This process is appealing in that it does not require us to take a stance on which ratings are accurate. Thus, we do not need to overweight some opinions or suppress others. Upon completing this process, we find that, as predicted by our theory, our newly computed aggregate ratings *i*) are less compressed than the ones displayed by the platform, and *ii*) better correlate with external measures of quality, such as the Academy Awards.

The rest of the paper is structured as follows: Section 2 surveys the literature; Section 3 presents our theoretical model; Section 4 describes the data and the empirical analysis; Section 5 concludes.

2 Related Literature

This paper adds to a large and highly multidisciplinary body of research studying both the nature and the effects of online consumer reviews. For overviews of this topic, see Cabral [2012] and Tadelis [2016].

A first strand of this literature has focused on quantifying the impact of ratings on choice. Seminal work by Chevalier and Mayzlin [2006], and more recent one by Luca [2016], among others, find sizable causal impact.

A second strand of research documents, theoretically and empirically, the nature of ratings, as well as their systematic biases. Systematic biases in ratings result from both sellers' strategic behavior (e.g. Chevalier et al. [2014], Nosko and Tadelis [2015] and Luca and Zervas [2016]) and as non-strategic equilibrium market outcomes (Li and Hitt [2008], Godes and Silva [2012], Brandes et al. [2013], Acemoglu et al. [2017], Besbes and Scarsini [2018], Bondi [2022]).

In the context of restaurants, Luca and Reshef [2021] show that ratings respond negatively to price increases. Because we are interested in isolating the role of consumers' quality reference points, we thus focus on a market with uniform pricing: movies (Orbach and Einav [2007]).

De Langhe et al. [2015] document low correlation between consumers and professional critics' opinions. Importantly, this holds even in markets without substantial product differentiation. Winer and Fader [2016] argue that low correlation is neither surprising nor necessarily problematic: less correlated sources of information are jointly more informative. This, however, is only true when the low correlation is due to, for instance, taste differences.

This paper, on the other hand, suggests it might be due to systematic biases in one of the two sources, complicating Winer and Fader [2016]'s conclusion.

A recent and fast growing literature has focused on platform and ratings design (Papanastasiou et al. [2018], Kremer et al. [2014]). These papers describe ways in which platforms can (benevolently) "distort" information to persuade (myopic) users to explore potentially promising options, instead of simply exploiting known ones. The platforms can achieve this goal by "spamming" new and unproven options and suppressing the ratings of the most popular ones. Our correction is largely opposite, in that average ratings will relatively reward lower quality options due to consumer self-selection.

Closely related to the ratings design section of our paper is Dai et al. [2012]. Like us, they advocate for a more sophisticated aggregation rule for individual opinions. In the context of restaurants, they find that most experienced reviewers have stricter standards, consistent with our model and data. However, they do not focus on the self-selection bias at the core of our paper.

The crucial assumptions in our paper – that satisfaction is relative, not absolute – has received extensive scrutiny across the social sciences for over four decades. Kanheman and Tverskey's celebrated model of loss aversion introduced the idea of a reference point around which outcomes are evaluated (see Kahneman and Tversky [2013] and O'Donoghue and Sprenger [2018] for thorough reviews). This reference points is left largely unspecified. We follow Kőszegi and Rabin [2006] in assuming that the reference point is the rational expectation over outcomes.[2]

We conclude this Section by reporting some non-academic references describing the movie rating process[3] – from consumers and critics alike – more specifically. On Reddit's "How do you rate movies" thread[4], one consumer argue that her ratings, rather than being absolute, *"[are] more [about] how they compare to the movies in the same genre"*. Another user describes a "mistakes spotting" technique: *"I don't have any sort of rubric. It's more based on how many faults I personally find in the movie."* Both are in line with the fact that experience negatively impact ratings.

Others make reference dependence even more apparent by stating they rate on curve: *"In reality, the majority of your movies will fall right in the middle - 3/5 (a normal bell curve). Only a handful out of 100 movies should ever get your 5/5 and a handful should get 0/5."*

[2]See also Bordalo et al. [2017] for a model of memory-based reference dependence. Note that in our setting the two formulations – backward and forward-looking – are largely equivalent, since the quality of past choices correlates (and, in our theoretical model, is in fact equal) to that of future ones.

[3]This process has a crucial impact of movies' commercial success, see for instance `https://www.theguardian.com/film/2015/sep/08/hollywood-rotten-tomatoes-jurassic-world-fantastic-four-terminator-genisys`.

[4]`https://www.reddit.com/r/movies/comments/aeb0ce/how_do_you_rate_movies/`

This quote from the late Roger Ebert[5] – arguably the most important movie critic of all time, and the only one to have been awarded the Pulitzer prize – is equally illuminating[6]:

> *"[T]he star rating system is relative, not absolute. When you ask a friend if "Hellboy" is any good, you're not asking if it's any good compared to "Mystic River," you're asking if it's any good compared to "The Punisher." And my answer would be, on a scale of one to four, if "Superman" (1978) is four, then "Hellboy" is three and "The Punisher" is two. In the same way, if "American Beauty" gets four stars, then "Leland" clocks in at about two."*

While the type of reference-dependence he suggests is subtler than the one in this paper, the two are consistent. The more consumers gain experience, the more they update their favorites in every genre, and then rate other movies in relation to them. By drawing from more movies, Experts form higher standards and post lower ratings.

3 The Model

We now present a simple theoretical model to organize our main assumptions and results. The model has two key primitives: consumers have heterogeneous abilities to screen quality, and, following Kőszegi and Rabin [2006], utility is reference-dependent, with the reference point uniquely pinned down by expectations.

There is a continuum of consumers divided in two types, Experts and Non-Experts, totaling mass 1. We denote by $\psi \in (0,1)$ the share of Experts. They choose between a continuum of vertically differentiated products, with quality $q \in [0,1]$.

Both types choose exactly one product - that is, the outside option is 0 and hence never chosen. In this sense, our model differs from a majority of theoretical work on the impact of online reviews, which focuses on whether reviews persuade consumers to buy a product over an outside option, rather than on which products are advantaged over their competing alternatives by reviews.

Experts choose according to $F_E(q)$, with density $f_E(q)$, Non-Experts $F_{NE}(q)$ ($f_{NE}(q)$). We make the following

Assumption 1 (Experience and Choice)**.** *On average, Experts identify and purchase better products, in a first order stochastic dominance sense. That is, for every $q \in [0,1]$, we have $F_E(q) \leq F_{NE}(q)$. Moreover, the two choice densities satisfy the MLRP property:*

$$\frac{\partial\left(\frac{f_E(q)}{f_{NE}(q)}\right)}{\partial q} > 0.$$

[5] `https://www.rogerebert.com/reviews/shaolin-soccer-2004`

[6] Relatedly, see `https://www.denofgeek.com/movies/are-star-ratings-on-movie-reviews-a-good-thing/` on how star ratings are contextual.

This assumption guarantees that Experts represent a larger share of buyers the higher the product's quality. For instance, Experts could observe more precise signals of quality, possibly due to access to better information sources, or have greater ability to interpret information.

Next, we model standards, or reference points. We assume that for each type of consumers, standards are defined as the expected level of quality, given choice procedures:

$$r_E := \int_0^1 q dF_E(q), \qquad r_{NE} := \int_0^1 q dF_{NE}(q).$$

Within this particular framework, because $F_E(\cdot)$ and $F_{NE}(\cdot)$ are fixed, one can think of r_E and r_{NE} as both expectations over the quality of future purchases and habits formed from previous choices.

It follows straightforwardly from the fact that $F_E(\cdot)$ first order stochastically dominates $F_{NE}(\cdot)$ that Experts form higher standards:

$$r_E = \int_0^1 q dF_E(q) > \int_0^1 q dF_{NE}(q) = r_{NE}.$$

We denote this gap by $\Delta(r) := r_E - r_{NE} > 0$.

Standards matter in shaping utility. Following Kőszegi and Rabin [2006], we make the following

Assumption 2 (Reference-Dependence)**.** *For every $q \in [0,1]$ and consumer type $i = E, NE$, we have*

$$U_i(q) = u(q) + \mu\big(u(q) - u(r_i)\big),$$

with $u(\cdot)$ satisfying the standard assumptions $u'(\cdot) > 0$ and $u''(\cdot) < 0$, and $\mu(\cdot)$ representing the classic Kahneman and Tversky [2013] gains-losses weight function: $\mu'(\cdot) > 0$, $\mu''(x) < 0$ if and only if $x > 0$.

Standards enter utility negatively. This leads Experts to be less satisfied than Non-Experts, for any level of q:

$$\begin{aligned} r_E > r_{NE} \Rightarrow & U_E(q) - U_{NE}(q) \\ & = \Big(u(q) + \mu\big(u(q) - u(r_E)\big)\Big) - \Big(u(q) + \mu\big(u(q) - u(r_{NE})\big)\Big) \\ & = \mu(u(q) - u(r_E)) - \mu(u(q) - u(r_{NE})) \\ & < 0 \quad \forall q \geq 0, \end{aligned}$$

where the inequality follows from $\mu'(\cdot) > 0$ and $u(q) - u(r_E) < u(q) - u(r_{NE})$.

We further assume that this gap in satisfaction translates into one in rating behavior:

Assumption 3 ((Subjectively) Honest Ratings)**.** *For every $q \in [0,1]$, and $i = E, NE$, ratings reflect subjective satisfaction:*

$$\mathcal{R}_i(q) = U_i(q).$$

Without loss of generality, given Experts' higher stringency, we can normalize utilities so that $\mathcal{R}_E(0) = 0$ and $\mathcal{R}_{NE}(1) = 1$. This ensures all ratings lie within this interval.

We first consider the case in which the average ratings displayed by the platform are the average of individual opinions. That is,

$$\begin{aligned}
\mathcal{R}(q) &= \frac{\psi f_E(q)\mathcal{R}_E(q) + (1-\psi) f_{NE}(q)\mathcal{R}_{NE}(q)}{\psi f_E(q) + (1-\psi) f_{NE}(q)} \\
&= \frac{\psi f_E(q)U_E(q) + (1-\psi) f_{NE}(q)U_{NE}(q)}{\psi f_E(q) + (1-\psi) f_{NE}(q)} \\
&=: \omega_E(q,\psi)U_E(q) + \big(1-\omega_E(q,\psi)\big)U_{NE}(q) \\
&= \big(1+\mu\big)u(q) - \mu\big(\omega_E(q,\psi)r_E + (1-\omega_E(q,\psi))r_{NE}\big),
\end{aligned}$$

where

$$\omega_E(q,\psi) := \frac{\psi f_E(q)}{\psi f_E(q) + (1-\psi) f_{NE}(q)}$$

represents the share of buyers who are Experts, as a function of product quality q and their baseline share ψ.

We will relax this assumption and consider aggregation rules that prioritize more experienced consumers, as often found in real world applications. There are obvious rationales for weighting some opinions more than others; however, we will show that this need not help with our bias.

We can now state our central result.

Proposition 1. *Average ratings understate quality differences. Moreover, ratings can be non-monotonic in quality. In particular, $\mathcal{R}'(q) > 0$ if and only if the following condition is satisfied:*

$$u'(q) \geq \frac{\partial \omega_E(q,\psi)}{\partial q} \cdot \Delta(r) \cdot \frac{\mu}{1+\mu}. \tag{1}$$

Proof. The proof for this and all other theoretical results can be found in the Appendix. ■

The first result follows directly from the fact that higher quality products are purchased by a higher share of Experts. Thus, they face a higher burden of proof, and their relative ratings are penalized.

The second result, while seemingly more complicated, admits nice intuition. To this end, it is useful to label each of the terms of Equation 1:

$$\underbrace{u'(q)}_{\text{Gains in individual satisfaction}} \geq \underbrace{\frac{\partial \omega_E(q,\psi)}{\partial q}}_{\text{Increase in \% of E buyers}} \cdot \underbrace{\Delta(r)}_{\text{Difference in standards}} \cdot \underbrace{\frac{\mu}{1+\mu}}_{\text{Importance of reference-dependence}}$$

In slightly greater detail, the LHS quantifies the gains in ratings from improved quality: each individual consumer is more satisfied, as measured by $u'(q)$. The RHS quantifies its costs, driven by self-selection: the first term represents choice heterogeneity, and the second and third rating heterogeneity. In particular, $\frac{\partial \omega_E(q,\psi)}{\partial q}$ represents negative self-selection, or the increase in the share of Experts buyers as q increases; $\Delta(r)$ represents the difference in standards between Experts and Non-Experts; $\frac{\mu}{1+\mu}$ measures the relative weight of reference-dependence in shaping total individual utility.

Note that when either $\frac{\partial \omega_E(q,\psi)}{\partial q}$, $\Delta(r)$ or μ go to zero, ratings are guaranteed to be increasing, since $u'(\cdot) > 0$. It is straightforward to see why: the first case corresponds to a lack of self-selection, the second to equal standards (and thus ratings) for the two categories, and the third to reference-independent – and, thus, homogenous across consumers – utility. In our empirical application, we will make use of these facts to unbias aggregate ratings.

3.1 Remedies

We now slightly relax the nature of the aggregation rule, while maintaining linearity. Empirically, a vast majority of platforms overweight the opinions of their most experienced reviewers. While there are obvious rationales for doing so – for instance, more experienced consumers might be less likely to post fake ratings, or more thorough in their quality evaluations – how does this interact with the self-selection bias we focus on? The following Corollary answers this question.

Proposition 2. *When the percentage of Experts, ψ, is low, overweighting their opinions worsens the bias, making it more likely for ratings to become non monotone in quality.*

To gather some intuition for the result, note that the bias gets worse when the crowd of buyers is more heterogeneous: if 90% of buyers are Experts, for instance, then overweighting their opinions bring aggregate ratings closer to essentially only reflecting the (homoegenous) opinions of Experts, yielding monotonicity: $\mathcal{R}'(q) \approx (1+\mu)u'(q) > 0$. Conversely, if – say – only 10% were Experts, then increasing their share makes the set of opinions reflective of a more diverse crowd, and thus aggregate ratings more likely to be non-monotone.

We conclude this Section by discussing another common feature of platforms' aggregation of ratings, meaning rewarding the ratings of products that are purchased by more consumers. In our setting, the number of buyers for a given product is proportional to the share of

Experts buyers: while everybody responds to quality, Experts do so more than Non-Experts. Therefore, more popular products are effectively facing a higher burden of proof. We thus provide an additional rationale for the platform rewarding products receiving more ratings, even in a setting in which a greater number of reviews does not bring more accuracy or credibility *per se*.

Proposition 3. *Denote by* $\mathcal{N}(q) := \psi f_E(q) + (1-\psi) f_{NE}(q)$ *the number of ratings for product* q*, and by* $\beta(\cdot)$ *a reward fucntion, with* $\beta(\cdot), \beta'(\cdot) > 0$*. Then, substituting average ratings* $\mathcal{R}(q)$ *with mass inflated ratings* $\beta(\mathcal{N}(q)) \cdot \mathcal{R}(q)$ *reduces ratings' contraction and improves monotonicity.*

4 Data and Empirical Strategy

This Section presents empirical findings in line with how we model the rating behavior by Experts and Non-Experts. We start presenting the dataset, obtained by combining data from two online platforms. Then, we provide suggestive evidence that 1) Experts tend to select better movies (choice heterogeneity) and 2) post lower ratings than Non-Experts (rating heterogeneity). Finally, we exploit the information about the fully rating history of users to shut down the different sources of bias presented in Proposition 1. Doing so, we can compute new normalized ratings that take into account the different stringency levels of users.

4.1 The Dataset

The dataset contains information about movies and movie ratings from two different online recommendation systems: MovieLens, and IMDb. MovieLens is an online platform launched in 1997 and run by GroupLens, a research group of the Department of Computer Science and Engineering at the University of Minnesota. It allows users to rate movies and received movie recommendations based on their ratings. We use the "MovieLens 25M" Dataset, publicly provided by GroupLens. It contains information about 25 million users' ratings displayed between January 1995 and November 2019.[7] We restrict our analysis to 9,448 movies that were rated by at least 30 users and produced after 1994. For this subset of movies, we merge information displayed on IMDb, one of the major online databases and recommendation systems related to movies. Like MovieLens, IMDb users can rate movies. Moreover, IMDb provides detailed information about movies' characteristics about the population of users who rate each movie. Accordingly, for each movie, we have the following data: release year;

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

genre; the number of AMPAS Academy nominations and awards;[8] the average ratings posted on MovieLens and on IMDb; and the total number of users who rated the movie on each platform. Regarding the IMDb users: we know the proportion of users between 18, and 29; between 30 and 44; and over 45 years old; the proportion of female users, and users from the US who rated each movie.

To identify expert users, we exploit the IMDb information about the Top1000 users. These are the 1,000 users "who have voted for the most titles on the webpage."[9] IMDb does not disclose the identity of these users and the number of movies rated by Top1000. Accordingly, users are unaware of their role, ruling out socially-driven explanations behind their rating behavior.[10] In the remaining part of the paper, we consider the Top1000 users Experts and we show how their rating behavior differs from all other users.

Table 1 presents some descriptive statistics about the sample of movies with no missing observations (present on MovieLens and IMDb). It includes information about movies' characteristics, movie ratings on IMDb and MoiveLens, and the information about IMDb users. The genre of almost 70% of movies is either Action, Comedy, or Drama. More than 4% of movies are nominated for the Academy awards and less than 2% are awarded. The distributions of ratings on IMDb and MovieLens are comparable. On IMDb, the average rating is 6.5 stars over 10 (with almost a one-star standard deviation). On MovieLens, 3.2 stars over 5 (with a half-star standard deviation). Moreover, on both platforms, movies are rated on average by thousands of users. IMDb Top1000 users post lower ratings relative to the Non-Expert users. Yet, the Top1000 ratings' dispersion has a similar size to the one for Non-Experts. Finally, 60% of IMDb users are between 30 and 44 years old, they are predominantly male, and only 30% percent are based in the US.

4.2 Choice Heterogeneity

Experts differ from Non-Experts since they are more capable of choosing high-quality products. Thus, we expect Top1000 users to watch and rate more high-quality movies. The main estimating equation to study this assumption is the following:

$$n_i^{Top1000} = \alpha + \beta_1 q_i + \beta_2 X_i + \epsilon_i, \tag{2}$$

[8] We restrict our attention to the "big five" awards (Best Actor, Best Actress, Best Direction, Best Picture, Best Writing) and Best Animation, Best Documentary, and Best Foreign Movie to include movies' categories that are seldom selected for the other awards.

[9] For further information, see `https://help.imdb.com/article/imdb/track-movies-tv/who-are-the-top-1000-voters-how-do-i-know-if-i-m-one-of-them/GA9WG44Q76JS3H34?ref_=helpart_nav_15#`

[10] See *e.g.* Jacobsen [2015].

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

	Mean	SD	N	Min	Max
movie characteristics					
Movie Year	2007	6.8	9448	1995	2019
Genre: Action (%)	17	.	9448	0	1
Genre: Comedy (%)	26	.	9448	0	1
Genre: Drama (%)	26	.	9448	0	1
Academy Nominated (%)	4.7	.	9448	0	1
Academy Awarded (%)	1.3	.	9448	0	1
movie ratings					
$\bar{r}_i^{IMDb}$	6.5	.98	9448	1.4	9.5
n_i^{IMDb} (thousands)	68	1.5e+02	9448	.05	2564
$\bar{r}_i^{Top1000}$	5.9	.88	9448	1	9
$n_i^{Top1000}$	263.01	191.02	9448	0	928
$\bar{r}_i^{Movielens}$	3.2	.46	9448	.8548	4.483
$n_i^{Movielens}$ (thousands)	1.5684	4.3542	9448	.031	72.67
movie audience					
$prop_i^{18-29}$ (%)	14	.	9448	0	.6178
$prop_i^{30-44}$ (%)	60	.	9448	0	.826
$prop_i^{45}$ (%)	26	.	9448	0	.8025
$prop_i^{female}$ (%)	21	.	9448	0	.7958
$prop_i^{US}$ (%)	30	.	9448	0	.9475

Note: The table includes all scraped movies and present 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.

where $n_i^{Top1000}$ is the total number of ratings by Top1000 users for movie i;[11] q_i is the unobserved quality of movie i; X_i is a set of controls related to movie characteristics and its audience; and ϵ_i is an error term. Movie quality is not perfectly observable. In our preferred specifications, we proxy quality using two measures that do not depend on platforms' feedback: the AMPAS nominations and awards. In particular, we use two dummy variables that equal 1 if a movie has received at least one nomination, or at least one award, respectfully. Nominated and awarded movies are also more *popular*. They are distributed in many movie theaters with expensive marketing campaigns. To partially account for the correlation between quality and popularity, we control for the total number of ratings of movie i, n_i^{IMDb}.

Table 2 presents the results of our estimates. Movies with AMPAS nominations or awards are significantly more watched and rated by Top1000 users. Results are robust to different specifications with genre, year fixed effects, and other dimensions related to the audience of each movie. Having at least one nomination determines an increase of more than 26% in the number of Top1000 users, whereas having at least one award leads to an increase of more than 13%. We control for the total number of users rating each movie. Accordingly, we are showing that a larger proportion of Experts rates nominated and awarded movies. An alternative explanation of these positive effects may be related to similarities in tastes between expert users and the members of the Academy who select nominated and awarded movies. We reject this theory providing two arguments. Tastes by Top1000 users and Academy members do no seem to be fully aligned since the effect for the awarded movies is significantly smaller than the one for nominated movies. Moreover, we confirm similar results when we repeat the same analysis using IMDb and MovieLens ratings and rankings as proxies for movie quality. Appendix Tables 3 and 4 show that Top1000 users tend to watch and rate movies with higher ratings and better ranked on IMDb and MovieLens. In Appendix Figure 5, we show the distributions of $n_i^{Top1000}$ for nominated and awarded movies and for movies with low IMDb ratings. We confirm the large magnitude of the positive relationship between movie quality and the number of Experts watching and rating the movie.

4.3 Rating Heterogeneity

On top of selecting higher-quality movies, Experts have higher standards than Non-Experts. In line with this assumption, the summary statistics in Table 1 suggest that, on average, Top1000 users post lower ratings than all other users. To be clear, it is not that Experts like different movies from Non-Experts: the two categories' ratings are very high correlated (0.89). Rather, they like each movie less, with the average difference being larger than half star. Figure 1 reinforces this claim with a stark result. In the graph on the right, we plot ratings of Top1000 and Non-Top1000 users for the 9,448 movies in our sample. Although

[11] Appendix Figure 4 shows the distribution of $n_i^{Top1000}$ for all movies in our sample. No movies are rated by all 1000 Top1000 users. Accordingly, issues related to censoring bias are not relevant in our analysis.

Table 2: Top1000 Users Are More Likely to Rate Academy Nominated and Awarded Movies

	(1)	(2)	(3)	(4)	(5)	(6)
Academy Nominated	75.82***	76.24***	74.72***			
	(5.999)	(5.993)	(5.634)			
Academy Awarded				40.40***	40.15***	42.13***
				(11.25)	(11.24)	(10.57)
n_i^{IMDb} (thousands)	0.888***	0.890***	0.839***	0.910***	0.913***	0.861***
	(0.00894)	(0.00895)	(0.00891)	(0.00898)	(0.00899)	(0.00896)
$prop_i^{female}$			43.25***			42.70***
			(11.85)			(11.95)
$prop_i^{18-29}$			124.6***			121.2***
			(18.03)			(18.18)
$prop_i^{30-44}$			558.4***			559.7***
			(16.40)			(16.54)
$prop_i^{>45}$			0			0
			(.)			(.)
$prop_i^{US}$			5.199			2.482
			(9.848)			(9.930)
Constant	198.7***	198.6***	-161.9***	200.3***	200.1***	-159.8***
	(1.355)	(1.353)	(11.85)	(1.360)	(1.358)	(11.95)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.6189	0.6215	0.6660	0.6130	0.6155	0.6603
N	9,448	9,448	9,448	9,448	9,448	9,448

Note: The outcome variable is the total number of ratings posted by Top1000 users. The sample includes all movies. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

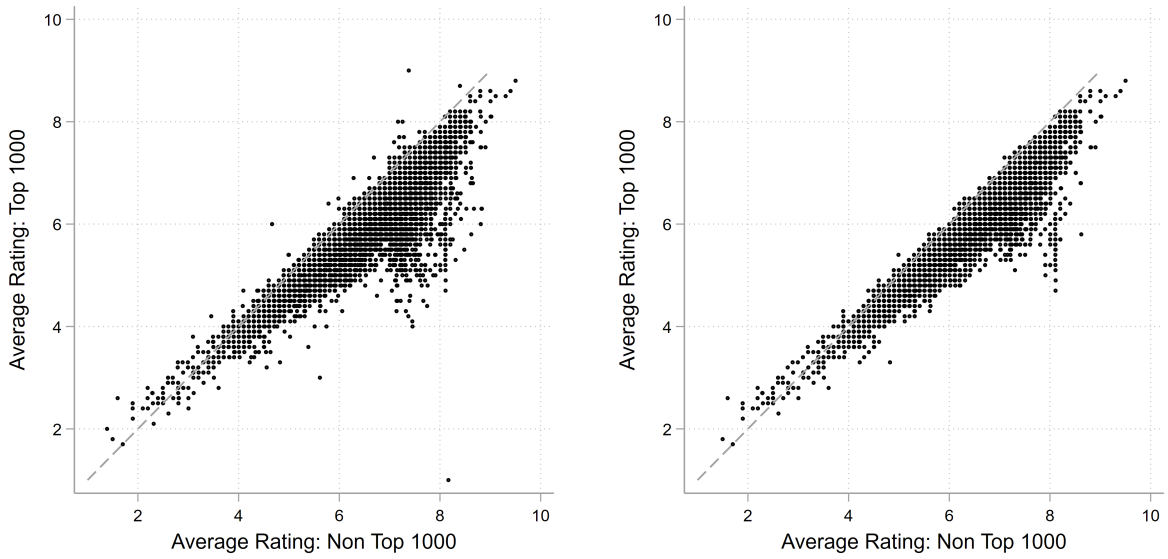


Figure 1: Average Ratings for Top1000 and Non-Top1000 Users: All Movies (right) and Only Movies Rated by at least 100 Top1000 Users (left).

Experts' and Non-Experts' ratings highly correlate, Top1000 users post lower ratings for more than 98% of the movies.[12]

The positive correlation between users' expertise and rating stringency is not limited to IMDb. We can exploit data about the total number of ratings posted by Movielens users to provide suggestive evidence also in this setting. In Appendix Figure 6 we compare the average ratings for MovieLens users with a different total number of ratings on the platform. Users with more ratings - thus more experienced - are more stringent and post lower ratings.

We conclude this Section by comparing ratings for movies with different quality levels.

In Figure 2, we show the average ratings by Top1000 and Non-Top1000 users for not-nominated, nominated, and awarded movies. Top1000 users always post lower ratings. Yet, this is particularly true for movies that do not receive Academy nominations. This is in line with the classical literature about reference points (Kahneman and Tversky, 2013) in which disappointment causes a larger loss in users' utility compared to positive surprises.[13]

Since Top1000 users have higher reference points, they are more likely to get disappointed than Non-Top1000 users. In particular, the differential between their ratings and ratings by Non-Expert increases for low-quality, not-nominated movies. We interpret this result as

[12]The number of Top1000 users who watch and rate movies is relatively small. To avoid potential biases, we repeat the same analysis only for movies that are rated by at least 100 Top1000 users in left graph of Figure 1. Again, Top1000 users post lower ratings for more than 98% of the movies.

[13]More technically, the gains-losses weight function is steeper and more convex for losses than it is for gains.

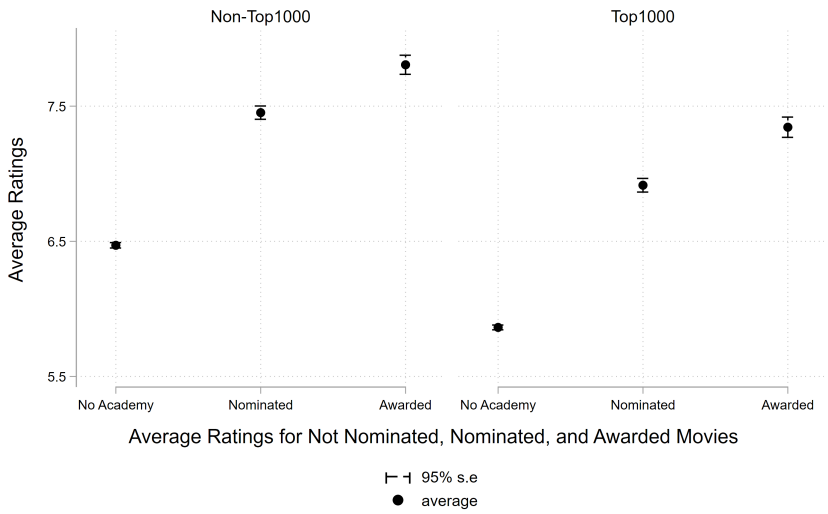


Figure 2: Top1000 Users Post Lower Ratings to Not Nominated or Awarded Movies

further indirect evidence about the relevance of reference points and users' loss aversion.

4.4 Debiasing Ratings

To reduce the bias due to users' heterogeneity in expertise, we propose to scale up ratings by exploiting users' rating history (in the language of Proposition 1, we equate the reference points for Experts and Non-Experts). This approach reduces the compression in movie quality by rating systems.

In the MovieLens dataset, we have access to the full rating history for each user and each movie. Accordingly, for each user j, it is possible to determine the average rating $\bar{r}_j$. Similarly, we can calculate the average rating $\bar{r}_i$ for each movie i. To determine the stringency of each user, we play with $\bar{r}_j$ and the average rating received by each movie rated by user j, denoted with $\bar{\bar{r}}_j$. In particular, the stringency of user j is $s_j = \bar{\bar{r}}_j - \bar{r}_j$. If user j tends to post scores that are lower (higher) than the average given by all the other users, then s_j is positive (negative) and user j can be classified as stringent. Then, we normalize ratings taking into account the stringency level of each user: $r_{ij}^{norm} = r_{ij} + s_j$. Finally, for each movie i, we calculate the new average normalized rating $\bar{r}_i^{norm}$.

This approach should reduce the bias related to user-specific reference points. However, our correction does not completely consider the self-selection between experts and high-quality movies. User are more or less stringent depending on the other users who rated the movies they select. Therefore, if experts exclusively rate a certain set of movies, they could appear less stringent they they actually are. Accordingly, the gains from our correction are only a lower-bound and do not completely eliminate the bias related to the heterogeneity in

users' experience.

Using our approach, we provide empirical evidence that *i*) more experienced users tend to be more stringent; *ii*) high-quality movies are rated by a more stringent audience; and *iii*) the bias against higher-quality products can be reduced by taking into consideration users' stringency. Figure 3b compares the stringency of MovieLens users with a different total number of ratings on the platform. Users with more ratings, thus more experienced, are more stringent. In Figure 3c, we present the average user stringency for not-nominated and nominated movies. Nominated movies face slightly more stringent users than not-nominated movies. Because of that, their ratings are downward-biased and high-quality movies are penalized. To show this, in Figure 3d, we show the distribution of original ratings ($\bar{r}_i$) and the normalized ratings ($\bar{r}_i^{norm}$) for nominated movies. The normalized ratings magnify the scores for nominated movies, confirming the negative bias present in the original ratings for high-quality movies.

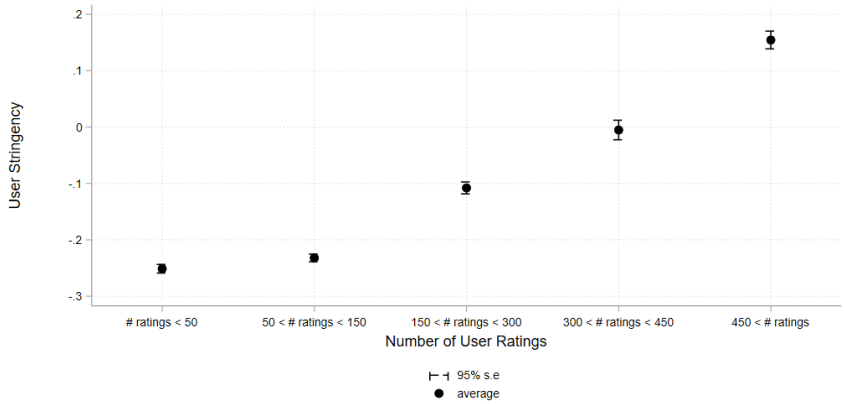
The IMDb dataset does not provide information about individual user ratings. Thus, we cannot adopt the strategy used in the previous context. However, we can try to ease rating aggregation by using the average ratings that Experts and Non-Experts assign to movies with similar characteristics. In particular, we can calculate the average ratings that Top1000 and Non-Top1000 users assign to movies of the same genre and produced in the same year. Then, we can subtract these aggregate averages from each movie's average for different user groups. In Appendix Figure 7, we plot these "debiased" ratings of Top1000 and Non-Top1000 users for all movies in our sample. As before, Experts' and Non-Experts' ratings are highly correlated. Yet, now Top1000 users do not post anymore lower ratings compared to Non-Top1000 users. Accordingly, this approach seems to facilitate the comparison of ratings by different categories of users.

5 Implications and Conclusion

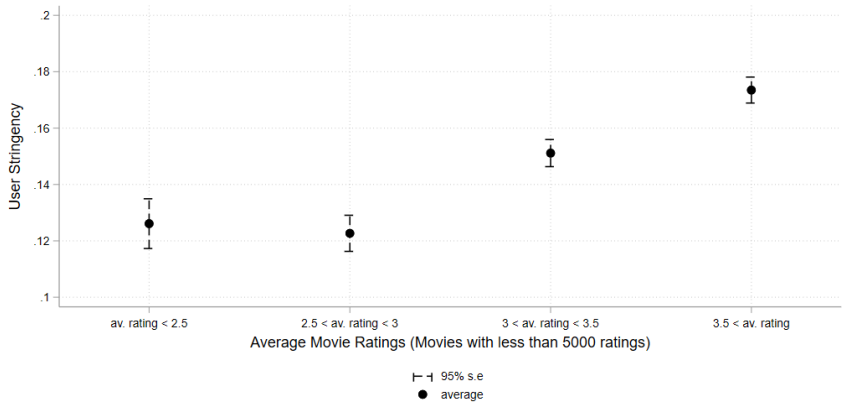
In this paper, we investigate the consequences of choice heterogeneity and reference-dependent preferences on the nature of online consumer ratings. We argue that better products are systematically purchased by more knowledgeable consumers who – as a natural byproduct of their superior choices – expect more, and thus post more stringent evaluations. As a result, high quality products' relative ratings are lower compared to those of their inferior alternatives.

Movie ratings provide an ideal scenario to test our theoretical claims. Consumer ratings play an increasingly crucial role in shaping movies' success. Moreover, fixed prices lead ratings to more closely represent quality.

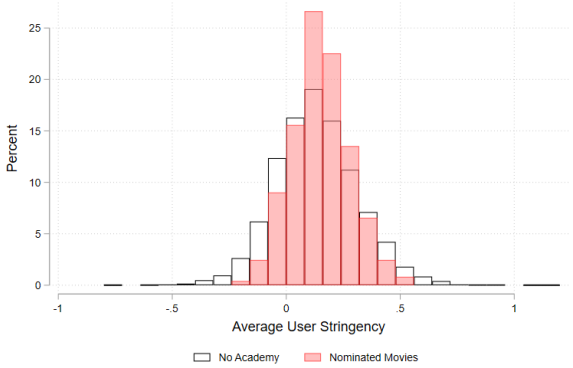
Using data from two large movie rating online platforms, we test our claims and find striking support for them. Experts rate higher quality movies, as proxied by Academy awards



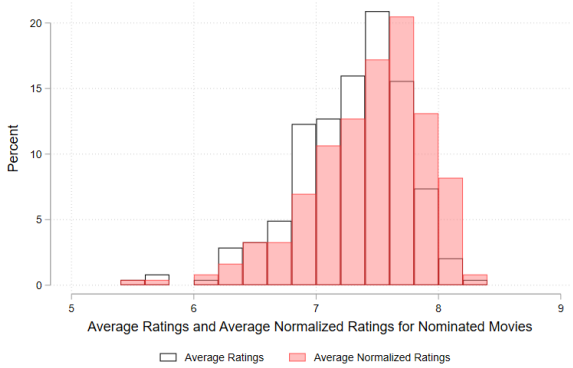
(a) Users' Stringency and Total Number of Ratings



(b) Users' Stringency and Average Ratings



(c) Users' Stringency for Nominated and Not-Nominated Movies



(d) Original and Normalized Ratings for Nominated Movies

Figure 3: Exploiting User-specific Information to Debias Ratings on MovieLens

wins and nominations. Moreover, Experts rate movies much more stringently. This is true for a striking 98.5% of movies in our sample. Combined, these two facts imply a bias against higher quality movies. We show how to correct for it theoretically and empirically, and debias ratings by exploiting the full history of users' ratings and equating their stringency levels. This approach leads to normalized aggregate ratings that better correlate with external proxies of quality, such as the Academy awards.

In thinking about the generalizability of our results, it is worth pointing out that, if anything, the movie market is fairly "egalitarian": certain movies (*e.g.* *"The Shawshank Redemption"* or *"Schindler's List"*) will be watched by a majority of both Experts and Non-Experts, and will thus contribute to shape standards for both consumers groups.

Thus, we believe that the bias we identify will be much stronger when looking at product categories (*e.g.*, restaurants) in which the discrepancy in choices, and thus standards, between Experts and Non-Experts is much more pronounced.[14] For some of these product categories, consumer ratings likely reflect an "apple-with-oranges" comparison, even more so than we show they do with movies.

Notably, 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. However, unlike our proposed solution, BeerAdvocate stops short of computing (and correcting for) product-specific stringency scores, which makes internalizing this information extremely difficult for consumers.

Last, we believe the mechanism we highlight to extend considerably more generally. One prominent example is US colleges grading policies. Moore et al. [2010] show that ratings standards vary considerably across colleges, and that employers do not properly correct for this bias, favoring students from more grade inflated institutions. This is despite the fact that the stakes for employers are much higher than those for those of most consumers, and colleges' grading policies are both transparent and widely debated.[15] For this reason, we believe this form of mislearning to be even more pervasive on online platforms, unfairly rewarding some products to the detriment of others.

[14]The main empirical disadvantage presented by most of these contexts, and the reason why we have not focused on them in our empirical analysis, is the key role played by prices, so that reviews need not solely represent quality. See Luca and Reshef [2021] for an empirical study of the impact of prices on restaurant ratings.

[15]For example, grade inflation in the Ivy League has received considerable media attention. Moreover, some universities, *e.g.* Princeton, have been proudly emphasizing their stricter grading standards compared to *e.g.* Harvard.

A Proofs

Proof of Proposition 1 To prove the first part, take $q_1 > q_2$ and note that without self-selection, we have

$$\hat{\mathcal{R}}(q_1) - \hat{\mathcal{R}}(q_2) = (1+\mu)\big(u(q_1) - u(q_2)\big), \quad \forall q_1 > q_2.$$

On the other hand, the average ratings actually observed on the platform are given by

$$\begin{aligned}
\mathcal{R}(q_1) - \big(1+\mu\big)u(q_1) &= -\mu\big(\omega_E(q_1,\psi)r_E + (1-\omega_E(q_1,\psi))r_{NE}\big) \\
&< -\mu\big(\omega_E(q_2,\psi)r_E + (1-\omega_E(q_2,\psi))r_{NE}\big) \\
&= \mathcal{R}(q_2) - \big(1+\mu\big)u(q_2),
\end{aligned}$$

where the inequality follows from the fact that $r_E > r_{NE}$ and $\omega_E(q_1,\psi) > \omega_E(q_2,\psi)$, for all $q_1 > q_2$ and $\psi > 0$. Rearranging the terms yields

$$\mathcal{R}(q_1) - \mathcal{R}(q_2) < \hat{\mathcal{R}}(q_1) - \hat{\mathcal{R}}(q_2),$$

proving the first part of the proposition. To prove the second part, note that

$$\mathcal{R}'(q) = (1+\mu)u'(q) - \mu\Big(\frac{\partial \omega_E(q,\psi)}{\partial q} r_E - \frac{\partial \omega_E(q,\psi)}{\partial q} r_{NE}\Big).$$

Rearranging terms, we have that

$$\mathcal{R}'(q) \geq 0 \Leftrightarrow (1+\mu)u(\prime q) \geq \mu \cdot \Delta(r) \cdot \frac{\partial \omega_E(q,\psi)}{\partial q},$$

as desired.

Proof of Proposition 2

Proof. First, note that overweighting experts by a factor $\gamma > 1$ means that now

$$\mathcal{R}(q) = \frac{\gamma\psi f_E(q)\mathcal{R}_E(q) + (1-\psi)f_{NE}(q)\mathcal{R}_{NE}(q)}{\gamma\psi f_E(q) + (1-\psi)f_{NE}(q)}.$$

This is the same rating that would be generated without overweighting, if experts were in proportion

$$\frac{\gamma\psi}{\gamma\psi + (1-\psi)} \geq \psi,$$

with equality only holding at $\psi = 0$ and $\psi = 1$. Thus, we can study this question in terms of an increase in ψ. Because ψ only enters Equation 1 through the term $\frac{\partial\omega_E(w,\psi)}{\partial q}$, to understand the impact of an increase in ψ we have to sign the second derivative, $\frac{\partial^2\omega_E(w,\psi)}{\partial q\partial\psi}$.

To simplify the rest of the proof, it is useful to start by proving the following

Lemma 1. *We can assume $f_{NE}(q) = 1$ for every q, without loss of generality.*

Proof. It is immediate to see that the transformation

$$\big(f_E(q),\ f_{NE}(q)\big) \to \left(\frac{f_E(q)}{f_{NE}(q)},\ 1\right)$$

leaves the proportion of experts buying each product – and thus $\mathcal{R}(q)$ – unchanged for every q. However, it needs not be the case that $\frac{f_E(q)}{f_{NE}(q)}$ integrates to 1 – that is, that it is an acceptable choice density function. To get around this problem, define

$$\alpha = \int_0^1 \frac{f_E(q)}{f_{NE}(q)} dq$$

and $\hat{f}_E(q)$ its normalised version, that is $\hat{f}_E(q) := \frac{f_E(q)}{f_{NE}(q)} \cdot \frac{1}{\alpha}$. Then, we can write

$$\begin{aligned}\omega_E(q) &= \frac{\psi\alpha\hat{f}_E(q)}{\psi\alpha\hat{f}_E(q) + (1-\psi)} \\ &= \frac{\frac{\psi\alpha}{\psi\alpha+(1-\psi)}\hat{f}_E(q)}{\frac{\psi\alpha}{\psi\alpha+(1-\psi)}\hat{f}_E(q) + \frac{(1-\psi)}{\psi\alpha+(1-\psi)}} \\ &= \frac{\psi'\hat{f}_E(q)}{\psi'\hat{f}_E(q) + (1-\psi')},\end{aligned}$$

where $\psi' := \frac{\psi\alpha}{\psi\alpha+(1-\psi)}$. Note that $\psi' \in [0,1]$ by construction, because $\alpha > 0$. Moreover, $\psi' = 0$ (resp. 1) when $\psi = 0$ (resp. 1), and ψ' is continuously increasing in ψ. It follows that this transformation, independently of α, does not restrict the set of admissible parameters. This completes the proof of the Lemma.

■

Now, assuming $f_{NE}(q) = 1$, we have $\omega_E(q,\psi) = \frac{\psi f_E(q)}{\psi f_E(q)+(1-\psi)}$, and thus, omitting some straightforward algebra,

$$\frac{\partial\omega_E(q,\psi)}{\partial q} = \frac{\psi(1-\psi)f'_E(q)}{\big(\psi f_E(q) + (1-\psi)\big)^2}$$

and

$$\frac{\partial^2\omega_E(q,\psi)}{\partial q\partial\psi} = \frac{\big(1-2\psi\big)f'_E(\psi f_E(q) + (1-\psi)\big) - 2(f_E(q)-1)\psi f_E(q)}{\big(\psi f_E(q) + (1-\psi)\big)^4}$$

Taking $\psi \to 0$, we have

$$\lim_{\psi\to 0}\frac{\partial^2\omega_E(q,\psi)}{\partial q\partial\psi} = \frac{\big(1-2\psi\big)f'_E(\psi f_E(q) + (1-\psi)\big) - 2(f_E(q)-1)\psi f_E(q)}{\big(\psi f_E(q) + (1-\psi)\big)^4} = f'_E(q) > 0,$$

where the last inequality comes from combining $f_{NE}(q) = 1$ and Assumption 1. By continuity in ψ, the expression remains positive whenever $\psi < \psi^*$, for some $\psi^* \in (0,1)$,

completing the proof.

■

Proof of Proposition 3

Proof. For every $q > 0$, we have that

$$\frac{\partial\Big(\beta(\mathcal{N}(q)) \cdot \mathcal{R}(q)\Big)}{\partial q} = \mathcal{N}'(q)\beta'(\mathcal{N}(q)) \cdot \mathcal{R}(q) + \beta(\mathcal{N}(q)) \cdot \mathcal{R}'(q)$$

We want to show that the RHS is positive whenever $\mathcal{R}'(q)$ is, or equivalently that $\mathcal{N}'(q)\beta'(\mathcal{N}(q)) \cdot \mathcal{R}(q) > 0$. Because $\mathcal{R}(q) \geq 0$ and so is $\beta'(\mathcal{N}(q))$, this boils down to showing $\mathcal{N}'(q) > 0$. But $\mathcal{N}(q) = \psi f_E(q) + (1-\psi) f_{NE}(q)$. Using Lemma 1, we can assume $f_{NE}(q) = 1$ without loss of generality. But then by the MLRP assumption $f'_E(q) > 0$, which completes the proof. ■

B Additional Figures and Tables

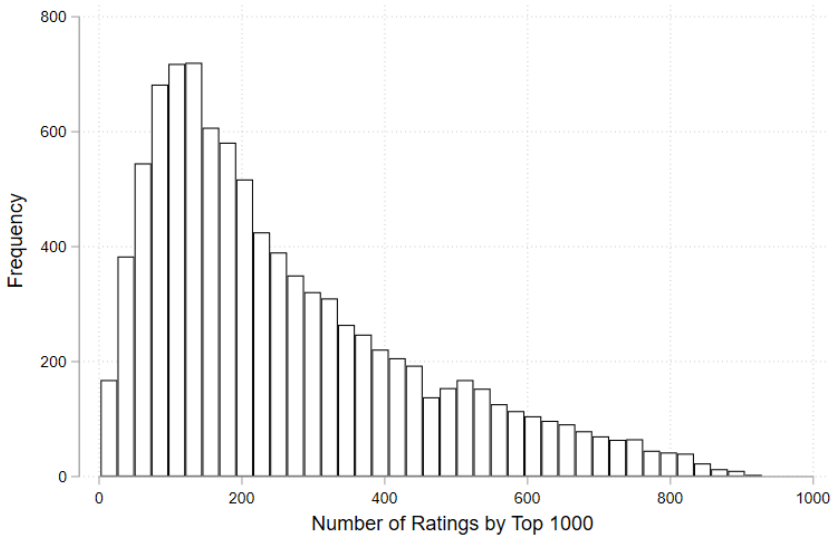


Figure 4: Distribution of the number of ratings posted by Top 1000 users.

Table 3: Top1000 Users Are More Likely to Rate Movies with Better IMDb Ratings and Ranking

	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{r}_i^{IMDb}$	43.80***	43.55***	34.36***			
	(1.962)	(1.968)	(1.891)			
$ranking_i^{IMDb}$				0.0160***	0.0159***	0.0118***
				(0.000712)	(0.000714)	(0.000695)
$prop_i^{female}$			9.562			19.73
			(17.00)			(17.15)
$prop_i^{18-29}$			625.7***			615.4***
			(24.71)			(24.93)
$prop_i^{30-44}$			759.5***			752.6***
			(23.22)			(23.27)
$prop_i^{>45}$			0			0
			(.)			(.)
$prop_i^{US}$			65.89***			60.39***
			(14.28)			(14.29)
Constant	-22.26*	-20.66	-525.8***	187.5***	188.0***	-352.5***
	(12.90)	(12.94)	(21.35)	(3.798)	(3.805)	(17.28)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.1945	0.1974	0.3211	0.1949	0.1977	0.3181
N	9,448	9,448	9,448	9,448	9,448	9,448

Note: The outcome variable is the total number of ratings posted by Top 1000 users. The sample includes all movies. Movie rankings are computed ordering movies using their ratings. A higher ranking value is associated with higher ratings (and higher quality). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Top1000 Users Are More Likely to Rate Movies with Better MovieLens Ratings and Ranking

	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{r}_i^{Movielens}$	80.53***	80.73***	68.18***			
	(4.152)	(4.178)	(3.941)			
$ranking_i^{Movielens}$				0.0140***	0.0140***	0.0109***
				(0.000704)	(0.000707)	(0.000672)
$prop_i^{female}$			-17.00			-7.113
			(16.90)			(16.98)
$prop_i^{18-29}$			658.3***			643.8***
			(24.52)			(24.70)
$prop_i^{30-44}$			765.8***			758.5***
			(23.27)			(23.31)
$prop_i^{>45}$			0			0
			(.)			(.)
$prop_i^{US}$			65.93***			58.62***
			(14.33)			(14.31)
Constant	2.234	1.593	-525.5***	197.0***	197.0***	-349.6***
	(13.56)	(13.65)	(21.78)	(3.770)	(3.784)	(17.31)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.1844	0.1878	0.3189	0.1859	0.1892	0.3162
N	9,448	9,448	9,448	9,448	9,448	9,448

Note: The outcome variable is the total number of ratings posted by Top 1000 users. The sample includes all movies. Movie rankings are computed ordering movies using their ratings. A higher ranking value is associated with higher ratings (and higher quality). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

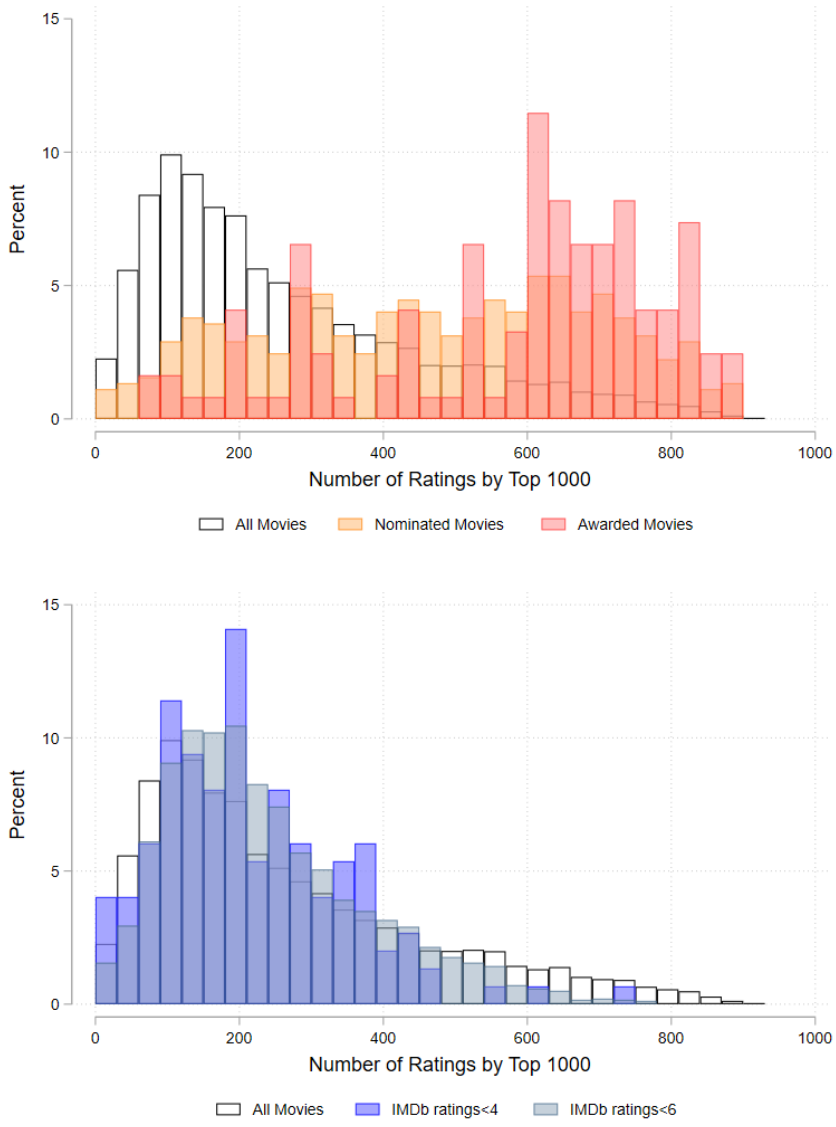


Figure 5: Distribution of the Number of Ratings Posted by Top1000 Users for Different Categories of Movies.

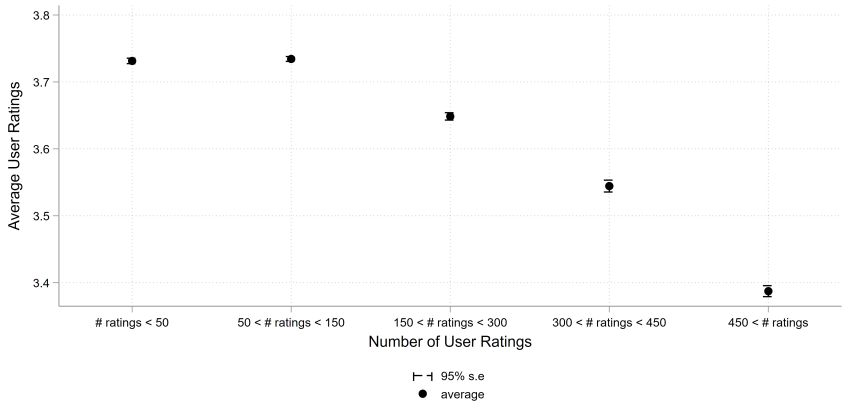


Figure 6: Average Ratings and the Total Number of Rated Movies on MovieLens.

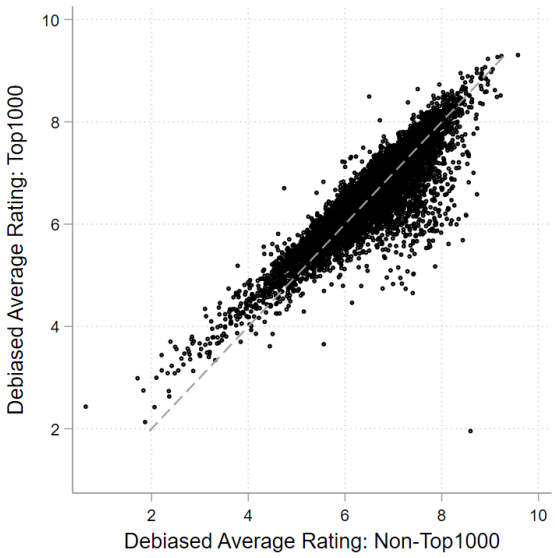


Figure 7: Exploiting User-specific Information to Debias Ratings on IMBd

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Fast and slow learning from reviews. Technical report, National Bureau of Economic Research, 2017.

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

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

Tommaso Bondi. *Alone, together*: A model of social (mis)learning from consumer reviews. 2022.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Memory, attention, and choice. *The Quarterly journal of economics*, 2017.

Leif Brandes, David Godes, and Dina Mayzlin. Controlling for self-selection bias in customer reviews. 2013.

Luis Cabral. Reputation on the internet. *The Oxford handbook of the digital economy*, pages 343–354, 2012.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

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

Weijia Dai, Ginger Z Jin, Jungmin Lee, and Michael Luca. Optimal aggregation of consumer ratings: an application to yelp. com. Technical report, National Bureau of Economic Research, 2012.

Bart De Langhe, Philip M Fernbach, and Donald R Lichtenstein. Navigating by the stars: investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, page ucv047, 2015.

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

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

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

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

Botond Kőszegi and Matthew Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, pages 1133–1165, 2006.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

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

Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016), 2016.

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

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

Don A Moore, Samuel A Swift, Zachariah S Sharek, and Francesca Gino. Correspondence bias in performance evaluation: Why grade inflation works. *Personality and Social Psychology Bulletin*, 36(6):843–852, 2010.

Chris Nosko and Steven Tadelis. The limits of reputation in platform markets: An empirical analysis and field experiment. 2015.

Ted O'Donoghue and Charles Sprenger. Reference-dependent preferences. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 1–77. Elsevier, 2018.

Barak Y Orbach and Liran Einav. Uniform prices for differentiated goods: The case of the movie-theater industry. *International Review of Law and Economics*, 27(2):129–153, 2007.

Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64(4):1727–1746, 2018.

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

Russell S Winer and Peter S Fader. 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, 2016.