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

Tommaso Bondi, Michelangelo Rossi and Ryan Stevens[*]

Cornell University, Télécom Paris, and Meta

October 24, 2022

Abstract

We study the impact of consumer quality-based self-selection on online reviews. Consumers differ in their expertise, which has two effects. First, expertise is instrumental to choice: Experts purchase better products than Non-Experts. Second, because of their superior choices, Experts endogenously form higher reference points, which leads them to 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: lower-quality products can obtain higher ratings than their superior alternatives, thanks to the lower standard they are held to. 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 festival and industry awards, we show that Experts rate movies of higher quality compared to Non-Experts. Moreover, and quite strikingly, Experts post more stringent ratings for approximately 98% of movies. Finally, we propose a fixed-point algorithm to debias the ratings. To do so, we exploit the full history of individual consumers' ratings to control for their individual stringency. This approach leads to normalized aggregate ratings that alleviate the bias against higher quality movies, and thus better correlate 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 ratings have become a ubiquitous driver of choice. But, to what extent can we trust their informational content? Because consumer ratings 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 expertise, which has two interrelated 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.

Therefore, the ratings of higher quality products reflect the higher standards of their buyers, and are thus downward biased relative to the ratings of their inferior alternatives. Simply put, ratings reflect an *"apples-with-oranges"* comparison, and in fact one that is particularly problematic, since one of their main roles is arguably that of allowing consumers to identify the highest quality options. Moreover, unlike taste-based self-selection, the quality-based self-selection mechanism we propose biases ratings comparisons even for products that are horizontally very similar.

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, price heterogeneity likely plays a key role in these two 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 expertise and expectation-based 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 separate lower and higher quality products. We can think of this as an adverse selection problem faced by sellers: the higher the quality produced, the less lenient the crowd purchasing and reviewing the product will be.

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 Expert consumers, the importance of standards in shaping utility, and the size of the utility gap due to expertise.

To elucidate the phenomenon, consider the following example: Expert and Non-Expert consumers decide between two movies, A and B. 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. Moreover, 90% of Experts choose movie A: their expertise is instrumental in identifying the superior option. Non-Experts, on the other hand, pick A and B randomly. 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 than those of Non-Experts for every movie. 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.[1]

We conclude our theoretical analysis by 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 expert consumers' 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, the most popular movie rating platform in the world[2], and complemented it with a massive (24 million reviews for 15,700 movies left by 162,000 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.

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

[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.

[2] As of 2022, IMDb has over 83 million registered users and is one of the top 50 most visited websites worldwide.

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 most relevant festival and industry awards around the world, and critics' reviews aggregated by Metacritic. Moreover, Experts are harsher than Non-Experts in their ratings. This is not only true on average, but the result holds for almost 98% of movies in our sample. This rating gap is statistically and economically significant: for a given movie on IMDb, Experts award, on average, over half-star less than Non-Experts.

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

The combination of Experts' quality-based self-selection and their stringent rating behavior implies that aggregate ratings penalize high quality movies compared to their inferior alternatives, as predicted by our theoretical model.

Last, we propose a fixed-point recursive algorithm to debias the ratings. The key observation to this end is that however pervasive, this type of self-selection bias allows for a straightforward correction. We exploit the full 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 these corrected individual ratings to re-compute all movie ratings. We iterate this process until it converges, that is, until individual stringencies and movie ratings are self-confirming.

This allows us to "level up the playing field", that is, to hold each movie to the same standard. Upon completing this process, we find that, as predicted by our theory, our debiased aggregate ratings better correlate with external measures of quality, such as nominations and awards, and reviews by critics. 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 more or less high quality.

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 informational content 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 ones by Anderson and Magruder [2012], Yoganarasimhan [2013], Luca [2016] and Farronato and Zervas [2022], 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 can result from both sellers' strategic behavior (e.g. Chevalier et al. [2014], Luca and Zervas [2016] and Carnehl et al. [2021b]), and various forms of consumer self-selection (Li and Hitt [2008], Godes and Silva [2012], Brandes et al. [2013], Acemoglu et al. [2022], Besbes and Scarsini [2018], Chung et al. [2020], Bondi [2022]).

In the context of restaurants, Luca and Reshef [2021] show that ratings respond negatively to price increases. Similarly, Carnehl et al. [2021a] use data from Airbnb and show that higher prices reduce the perceived value-for-money of Airbnb hosts resulting in lower ratings.

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]).

Speaking to our attempts to reconcile consumer-generated ratings with other, non-consumer-generated proxies of quality, 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 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, we argue, average ratings will endogenously relatively reward lower quality options.

Closely related to the ratings design section of our paper is Dai et al. [2012]. In the context of restaurants, they find that most experienced reviewers have stricter standards, consistent with our model and data. Similarly, Rocklage et al. [2021] finds that expertise numb consumer experiences. However, neither of these two papers focus on the self-selection bias at the core of our paper.

The crucial assumptions in our paper – that utility is relative, not absolute – has received extensive scrutiny across the social sciences for over four decades. Kanheman and Tversky's celebrated model of loss aversion introduced the idea of a reference point around which outcomes are evaluated. This reference point is left largely unspecified. We follow Kőszegi and Rabin [2006] in assuming that the reference point is the expectation over outcomes.[3]

Bushong and Gagnon-Bartsch [2016] propose a model in which as agents' reference points increase, their beliefs about quality become lower. They apply their model to study individual – but not social – learning, and confirm its main predictions experimentally. The fact that consumers are unaware of their increasing standards might help explaining why their ratings become harsher, and further suggests that peers' standards are likely not internalized when learning from ratings.[4]

Brandes et al. [2013] focus on the fact that selection happens at the ratings phase, conditional on choice: only those with extreme opinions about a product review it. We assume away self-selection at the rating stage, and only have it at the choice stage. This seems justified in our datasets, where a lot of Non-Extreme ratings are posted (for instance, we see a lot of 5s, 6s, 7s and 8s on a $1-10$ scale, and the histogram of ratings is bell-shaped). We believe that self-selection on choice is as (if not more) intuitive, sizeable and systematic as that on rating conditional on choice.

We conclude this Section by reporting some non-academic references describing the movie rating process – from consumers and critics alike – more specifically. On the Reddit's "How do you rate movies" thread,[5] one consumer argues 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 greater expertise 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."*

This quote from the late Roger Ebert[6] – arguably the most important movie critic of all

[3] See also Bordalo et al. [2017] for a model of memory-based reference dependence. In our theoretical setting the two formulations – backward- and forward-looking – are largely equivalent, since for each type of consumers the quality of choices is constant over time.

[4] Moreover, note that when learning from aggregate ratings, the identity of individual writers is unknown. Furthermore, even if consumers were fully aware that more expert peers tend to buy higher quality products and rate more strictly, product's qualities are unknown to them in the first place – a pre requirement for employing consumer reviews.

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

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

time, and the only one to have been awarded the Pulitzer prize – is equally illuminating[7]:

> *"[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."*

The more consumers gain expertise, 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 builds on two key primitives: first, consumers are heterogeneous in their expertise. More expert consumers have superior ability to select high quality products. Second, each consumer's utility is reference-dependent: it increases with the quality of the good purchased, and decreases with their expectation of quality.

More formally, there is a continuum of consumers totalling mass 1. We denote by $\psi \in (0,1)$ the share of Experts (E), by $1-\psi$ the share of Non-Experts (NE). Each consumer chooses between a continuum of vertically differentiated products, with quality $q \in [0,1]$. We normalize prices to 0.

Both types of consumers 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 self-selection biases of online reviews (such as Acemoglu et al. [2022]) in an important way. While existing models consider the monopolistic case, in which reviews can persuade consumers to buy a product over an outside option, we focus on which products are relatively advantaged over their competing alternatives by reviews. In other words, we focus on scenarios in which reviews help consumers decide *what*, not *if* to buy, which we believe are both ubiquitous and understudied.

Experts choose according to a continuous and smooth cumulative distribution $F_E(q)$, with density $f_E(q)$, Non-Experts $F_{NE}(q)$ ($f_{NE}(q)$). We make the following:

Assumption 1 (Expertise and Choice)**.** *On average, Experts identify and purchase better products. That is, for every $q \in [0,1]$, we have $F_E(q) \leq F_{NE}(q)$ (first order stochastic*

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

dominance). Moreover, the two choice densities satisfy the MLRP property:

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

These assumptions guarantee that *i*) Experts experience higher quality on average, and *ii*) the higher the product's quality, the higher the share of its buyers that are Experts. While presented as an Assumption here, this feature of the model allows for natural microfoundations. 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 individual consumers' 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, since $F_E(\cdot)$ and $F_{NE}(\cdot)$ are fixed over time, 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 $\Delta(r) := r_E - r_{NE} > 0$. In our empirical analysis, we find a substantial $\Delta(r)$ on both IMDb and MovieLens.

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*

$$U_i(q) = u(q) + \mu\big(u(q) - u(r_i)\big), \tag{1}$$

with $u(\cdot)$ satisfying the standard assumptions $u'(\cdot) > 0$ and $u''(\cdot) < 0$, and $\mu > 0$.[8]

It is apparent from Equation 1 that standards enter utility negatively, with μ quantifying

[8] This is a semplification. Formulating a more detailed model with $\mu(\cdot)$ being the classic loss averse function – that is, $\mu'(\cdot) > 0$, $\mu''(x) > 0$ if and only if $x < 0$ – yields similar insights, while making the exposition more cumbersome.

their impact. This leads Experts to be less satisfied than Non-Experts, for any level of q:

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

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

Our last, natural assumptions allows us to link differences in individual utilities to differences 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.[9]

We first consider the case in which the average ratings displayed by the platform are the average of individual opinions. That is, given a share ψ of Experts, choice densities $f_E(\cdot)$ and $f_{NE}(\cdot)$, and ratings $R_E(\cdot)$ and $R_{NE}(\cdot)$, we have:

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

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, their baseline share ψ, and choice densities $f_E(\cdot)$ and $f_{NE}(\cdot)$.

In Section 3.1, we relax this assumption and consider slightly more general aggregation rules, such as those that prioritize more expert consumers, and those that prioritize products

[9] Formally, this requires $u(0) = \frac{\mu \cdot u(r_E)}{1+\mu}$ and $u(1) = \frac{1+\mu \cdot u(r_{NE})}{1+\mu}$.

with a larger number of ratings, as commonly found in real world platforms (*e.g.*, Yelp, Amazon). 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. *Quality-based self-selection causes ratings to underestimate 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:*

$$u'(q) \geq \frac{\partial \omega_E(q,\psi)}{\partial q} \cdot \Delta(r) \cdot \frac{\mu}{1+\mu}. \tag{2}$$

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", which implies their relative ratings are penalized compared to those of their lower quality alternatives.

The second result, while seemingly more complicated, admits nice intuition. To this end, it is useful to label each of the terms of Equation 2:

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

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 types of consumers, and the third to reference-independent – and, thus, homogenous across consumers – utility.

In Section 4.4, we propose a fixed-point algorithm to back out individual stringency for each user, which allows us to correct for it and thus obtain new ratings reflecting $\Delta(r) = 0$, thus eliminating the bias.

3.1 Platform Design

We now relax the aggregation rule discussed in this Section so far, and consider two commonly observed real world practices. First, a vast majority of platforms overweight the opinions of their most expert reviewers. While there are obvious rationales for doing so – for instance, more expert consumers might be less likely to post fake ratings, or more thorough in their quality evaluations – the following Corollary shows that this can have perverse effect in our context.

Corollary 1. *When the percentage of Experts, ψ, is low, overweighting their opinions by a factor $\gamma > 1$ worsens the bias, further contracting ratings and thus making it more likely for the monotonicity between ratings and quality to be broken.*

The key observation towards gathering some intuition for the result is that the bias gets worse when the crowd of buyers is more heterogeneous. If 90% of buyers were Experts, for instance, then overweighting their opinions bring aggregate ratings closer to essentially only reflecting the (homogeneous) opinions of Experts, yielding monotonicity: $\mathcal{R}'(q) \approx (1 + \mu)u'(q) > 0$. Conversely, if – say – only 10% were Experts (which we believe to be the far more likely scenario empirically), increasing their share worsens the bias.

Interestingly, one of the platforms adopting this policy is exactly one of our focuses in this paper, IMDb. As IMDb consists of millions of casual movie watchers and a much smaller number of Experts, our results suggests its design is worsening the bias against high quality movies.

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*.

Corollary 2. *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 function, 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.*

We emphasize that our model describes a very stylized relationship between quality and number of ratings, in which higher quality products are on average more popular than their lower quality alternatives. While fairly natural, this relationship need not always hold empirically. In particular, one can think of products of mediocre quality but huge mass appeal, and viceversa, high quality and niche appeal. In those situations, the impact of inflating the ratings of popular products would be reversed.

In Section 4.4 we propose a complementary design remedy that is both new (to the best of our knowledge) and conceptually orthogonal to those highlighted in Corollaries 1 and 2. Unlike in Corollary 1, our algorithm counts all ratings equally, and thus does not require us to take a stance on which opinions are credible or not; it also requires less knowledge of the proportion of Experts (even more: it does not require us to take a stance on who is an expert in the first place). Moreover, unlike in Corollary 2, our algorithm is agnostic as to the predictive power of the number of ratings on ratings' stringency, and correct for the bias in either of the two cases highlighted here.

4 Data and Empirical Strategy

We now provide empirical support for the role of consumer expertise on both choice and rating behavior as described in Section 3, and study its consequences for movie ratings.

We start by presenting the dataset in Section 4.1. Our dataset is obtained by combining data from two online platforms: MovieLens and IMDb. In particular, we carefully explain how we proxy movies' quality with festival and industry awards, and users' stringency with the number of ratings posted on the platforms. Then, in Subsections 4.2 and 4.3, we provide empirical evidence for our two main assumptions. In particular, we will center our analysis on the following two empirical counterparts of Assumptions 1 and 2 in Section 3:

Choice Heterogeneity: *Experts watch and rate higher-quality movies than Non-Experts.*

Rating Heterogeneity: *For a given movie, Experts post lower ratings than Non-Experts.*

After providing empirical support for these two facts, in Subsection 4.4 we provide a fixed-point algorithm to remove the self-selection of stringent, expert reviewers into high quality movies. This technique allows us to compute new, normalized ratings, and provide convincing evidence for the main proposition of the model (Proposition 1): average ratings understate differences in quality, thus unfairly penalizing high quality movies compared to their inferior alternatives.

4.1 The Dataset

Our dataset combines information on movies, movie ratings and users 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 receive 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.[10]

We restrict our analysis to the 9,426 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. The two datasets exhibit nice complementarities. MovieLens provides individual ratings for users who rated at least 20 movies. Being able to identify the rating history for each user will be key to accounting for users' stringency and our algorithm to debias ratings in Subsection 4.4.

Conversely, IMDb provides only aggregate ratings, but detailed information about movies' characteristics and the population of users who rate each movie. For each movie, we have the following data: release year; runtime; genre; the number of nominations and awards to the major festival and industry awards around the world; the average ratings posted on MovieLens and on IMDb; the total number of users who rated the movie on each platform; whether a movie is reviewed by movie critics selected by Metacritic; and the aggregated Metacritic ratings based on critics reviews.[11] The number of nominations and awards and the Metacritic ratings are measures of quality that are external to the review systems. We will use them to show that high quality movies are watched and rated by a different crowd of users and this selection negatively affects their ratings.

Regarding IMDb users, we know, for each movie, 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."[12] IMDb does not disclose the identity of these users or the number of movies each of them has rated. This implies that users are unaware of their role, ruling out socially-driven explanations behind their rating behavior.[13] We use the Top1000 users as Experts and we show how their choice and rating stringency differ from all other users.

Table 1 presents some descriptive statistics about the 9,426 movies in the sample such as the movie genre, runtime, the year of production, and the characteristics of users who rated them. It includes information about movies' characteristics, movie ratings on IMDb, MovieLens, and Metacritic; and information about IMDb users. The genre of almost 70% of movies is either Action, Comedy, or Drama. Almost 60% of movies have been nominated or awarded by at least one festival or industry award. If we exclude the awards by the Academy

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

[11] Metacritic is a website that aggregates critics' reviews for movies, tv series, and videogames. For more information about Metacritic and they way they select movie critics' reviews, see `https://www.metacritic.com/about-metascores`.

[12] 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#`

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

of Motion Picture Arts and Sciences (better known as Oscar), the proportion of nominated or awarded movies goes down to less than 50%. 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, the average movie in our sample is rated by several thousands of users. 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.

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	104.1	28.9	9426	2	629
Genre: Action (%)	17	.	9426	.	.
Genre: Comedy (%)	26	.	9426	.	.
Genre: Drama (%)	26	.	9426	.	.
Nominated or Awarded (%)	58	.	9426	.	.
Nominated or Awarded (Excluding Academy) (%)	48	.	9426	.	.
movie ratings					
IMDb Ratings: $\bar{r}_i^{IMDb}$	6.5	.979	9426	1.4	9.5
IMDb Number of Users: $n_i^{IMDb}(\times 1000)$	69	148	9426	.05	2588
IMDb Top1000 Ratings: $\bar{r}_i^{Top1000}$	6	.864	9426	1	9
IMDb Number of Top1000 Users: $n_i^{Top1000}$	260.3	191	9426	2	928
MovieLens Ratings: $\bar{r}_i^{Movielens}$	3.2	.459	9426	.8548	4.483
MovieLens Number of Users: $n_i^{Movielens}(\times 1000)$	1.6	4.36	9426	.031	72.67
Have Critics Reviews: $Meta_i$ (%)	72	.	9426	.	.
Have Positive Critics Reviews: $Meta_i^{>60}$ (%)	32	.	9426	.	.
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 9426 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.

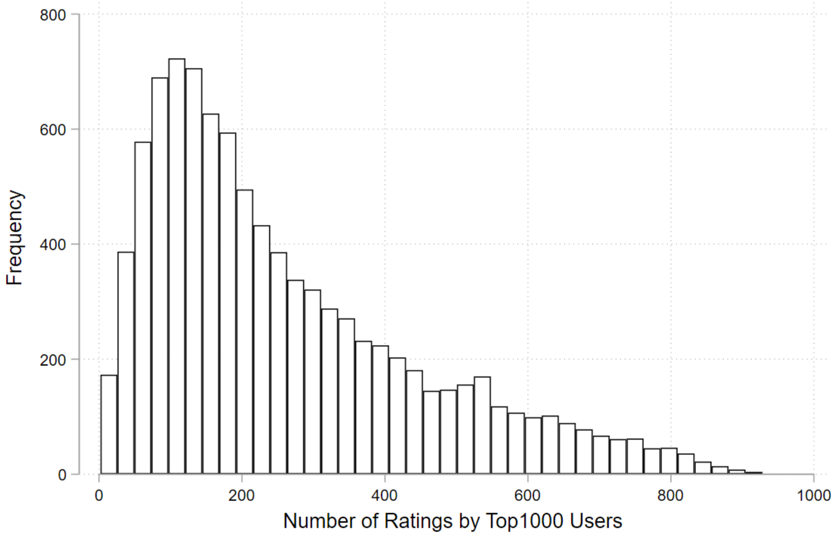


Figure 1: Distribution of the number of ratings posted by Top 1000 users.

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:

$$n_i^{Top1000} = \alpha + \beta_1 q_i + \beta_2 X_i + \epsilon_i, \tag{3}$$

where $n_i^{Top1000}$ is the total number of ratings by Top1000 users for movie i; Figure 1 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. 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 three measures that do not depend on platforms' feedback. In particular, we use one dummy variable that equals 1 if a movie has received at least one nomination or award to one of the major festival or industry awards around the world. A second dummy variable that equals 1 if a movie has received at least one nomination or award excluding the Academy awards. We exclude the Academy awards since previous research has shown that the relevance of the Oscars is such that it may move users' reference point, and affect selection of users who watch the nominated and awarded movies (Rossi, 2021). Finally, we use a third dummy variable that equals 1 if a movie has received a Metacritic Metascore greater than 60.

Quality and popularity are correlated positively. For instance, nominated and awarded movies are distributed in many movie theaters with expensive marketing campaigns. Thus, we can expect them to attract more users irrespective of their expertise. To partially account for the correlation between quality and popularity, we control for the total number of ratings of movie i, n_i^{IMDb}. Moreover, our focus on awards that are less well-known than the Academy awards should partially reduce the positive correlation between quality and popularity.[14] Table 2 presents the results of our estimates. Movies with nominations or awards, and with positive Metacritic reviews are significantly more watched and rated by Top1000 users. Results are robust to different specifications with genre, year fixed effects, and other controls that are related to the audience of each movie.

To put this in perspective, consider two sets of movies with the same number of total ratings (therefore with the same overall popularity) and similar movie and audience characteristics. Results in Table 2 suggest that the movies with at least one nomination or award (excluding the Academy awards) are watched and rated by almost 5% more Top1000 users, whereas having at least one nomination or award (including the Academy) is associated with an increase of more than 15%. Similarly, receiving positive reviews by movie critics is associated to a 10% increase in the number of Top1000 ratings, for a fixed number of total ratings. In other words, Experts are more responsive to quality than Non-Experts, as in our theorizing. An alternative explanation of these positive effects may be related to similarities in tastes between expert users and critics, or members of the committees who select nominated and awarded movies. Yet, we can control for a variety of variables that may account for specific tastes and the results remain positive and significant. Furthermore, as we highlight in greater detail in Section 4.3, there is an almost perfect correlation between the scores of Experts and Non-Experts, which rules out the possibility that Experts simply like a different set of movies. In Appendix Figure 5, we show the distributions of $n_i^{Top1000}$ for nominated and awarded movies, and for movies with Metacritic reviews and with positivie Metacritic reviews. By doing so, 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

Next, we show that 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. In Figure 2a, we plot ratings of Top1000 and Non-Top1000 users for the 9,426 movies in our sample. Top1000 users post lower ratings for 98% of the movies. 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

[14] Among others, the festival and industry awards considered here are the Golden Globes, the Cannes Film Festival, or the Venice Film Festival. But also, the Boston Society of Film Critics Awards, the César Awards, the Florida Film Critics Circle Awards.

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

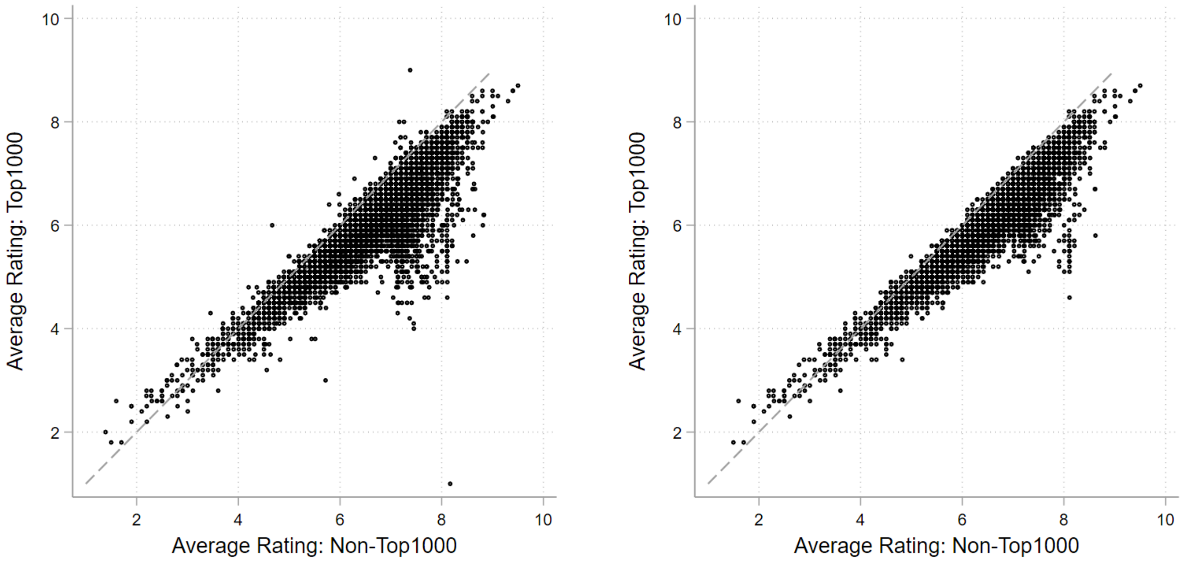
	(1)	(2)	(3)	(4)	(5)	(6)
Nominated or Awarded	42.38***	42.47***	38.51***			
	(2.75)	(2.75)	(2.62)			
Nominated or Awarded (Excluding Academy)				12.87***	13.36***	11.37***
				(2.51)	(2.51)	(2.37)
$Meta_i^{>60}$	18.32***	19.54***	27.01***	31.47***	32.62***	39.00***
	(2.92)	(2.93)	(2.77)	(2.82)	(2.83)	(2.66)
$n_i^{IMDb}(\times 1000)$	0.86***	0.86***	0.79***	0.88***	0.89***	0.81***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Movie Runtime			0.28***			0.32***
			(0.04)			(0.04)
p_i^{female}			46.73***			48.42***
			(11.62)			(11.74)
p_i^{18-29}			131.59***			136.61***
			(17.91)			(18.10)
p_i^{30-44}			573.04***			573.97***
			(15.93)			(16.10)
p_i^{US}			38.81***			21.23**
			(9.83)			(9.89)
Constant	170.74***	170.15***	-239.23***	183.04***	182.29***	-227.65***
	(1.89)	(1.90)	(12.58)	(1.85)	(1.85)	(12.73)
Genre FE	✓	✓	✓	✓	✓	✓
Year FE		✓	✓		✓	✓
R^2	0.630	0.633	0.680	0.622	0.625	0.674
N	9,426	9,426	9,426	9,426	9,426	9,426

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

100 Top1000 users in Figure 2b. Again, Top1000 users post lower ratings for approximately 98% of the movies. This rating gap is statistically and economically significant: Experts ratings are, on average, 0.57 stars lower than Non-Experts', for a given movie. This variation accounts for more than half of the standard deviation of movie ratings on IMDb.

To be sure, and again to reinforce the idea that systematic taste differences between the two groups are not driving our results, it is not the case that Experts like different movies than Non-Experts: Experts and Non-Experts ratings are very highly correlated (0.89). Rather, Experts rate nearly each movie lower, with the average difference being larger than half star.

Moreover, notice that this framework does not rule out individual taste differences, as long as *i*) the taste distribution is the same for Experts and Non-Experts and *ii*) there are enough Experts and Non-Experts rating each movie. The latter is achieved since the selected movies have, on average, more than 65,000 ratings on IMDb. As for the former, we believe our empirical results in the Appendix Figure 6 provide convincing support for it: we repeat the analysis for the four more common genres (Action, Comedy, Drama, and Other Genres) separately. Ratings by Experts and Non-Experts are highly correlated for each genre, suggesting a similar distribution of tastes between the two groups. Similarly, in Appendix Figure 7, we show the average ratings by Top1000 and Non-Top1000 users for



(a) All Movies

(b) Movies Rated by at least 100 Top1000

Figure 2: Average Ratings for Top1000 and Non-Top1000 Users

not nominated, and nominated or awarded movies (with and without the Academy awards). Again, in line with our model, Experts and Non-Experts respond to quality in very similar ways, but Top1000 users always post lower ratings.

Combined, choice and rating heterogeneity imply a bias against higher quality movies. To remove this bias, in Section 4.4 we illustrate a technique to correct for the stringency of each user. This allows us to quantify – and alleviate – consumer quality-based self-selection, improving the external validity of ratings.

4.4 Debiasing the Ratings

To debias the ratings, we exploit each user's MovieLens rating history to normalize their ratings' stringency (in the language of Proposition 1, we correct for the reference points, or standards, of Experts and Non-Experts). We then recompute aggregate ratings using the normalized individual ones.

Of course, individual stringency and ratings are intertwined: updating one implies updating the other. Thus, we iterate this process until it converges. More formally, for each MovieLens user j, defined as $\mathcal{I}_j$ the set of movies she has watched. Moreover, define by n_j the cardinality of $\mathcal{I}_j$, or user expertise. Then, for each movie $i \in \mathcal{I}_j$, compute the movie-user specific stringency as the average movie rating, $\bar{r}_i$, minus the rating posted by the user, r_{ij}. Then, we define the user's stringency as the average of movie-user specific stringency over the set $\mathcal{I}_j$:

$$s_j := \frac{\sum_{i \in \mathcal{I}_j} \left(\bar{r}_i - r_{ij}\right)}{n_j}. \tag{4}$$

Then, we compute new movie ratings which take into account – and correct for – how stringent their watchers are. Define by $\mathcal{J}_i$ 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

$$\bar{r}_i^{norm} := \frac{\sum_{j \in \mathcal{J}_i} \left(r_{ij} + s_j\right)}{n_i} = \bar{r}_i + \frac{\sum_{j \in \mathcal{J}_i} s_j}{n_i}. \tag{5}$$

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 consumers.

Equations 4 and 5 are clearly interdependent. How stringent each individual user is depends on how our normalization affects the ratings of the movies she has watched, and this normalization in turn depends on individual stringencies.

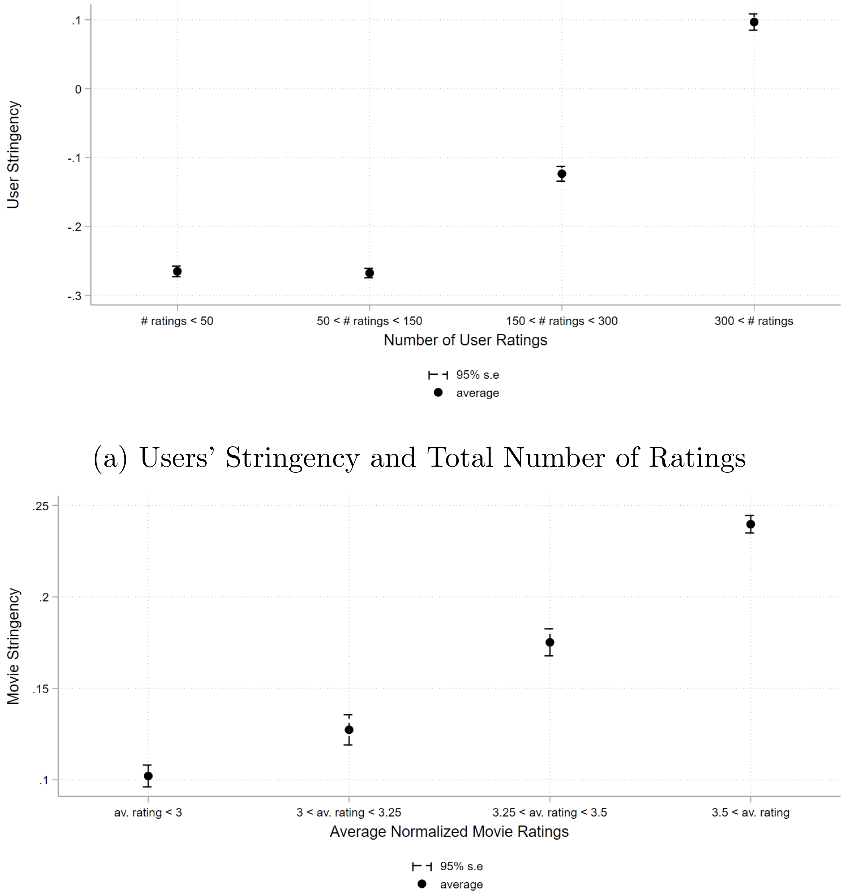
We solve this system of equations numerically. By doing so, we can define the normalized user stringency s_j^{norm} and the normalized movie stringency s_i^{norm} (or the average stringency of the movie audience) as

$$s_j^{norm} := \frac{\sum_{i \in \mathcal{I}_j} \left(\bar{r}_i^{norm} - r_{ij}\right)}{n_j}, \quad s_i^{norm} := \frac{\sum_{j \in \mathcal{J}_i} s_j^{norm}}{n_i}.$$

By construction, this approach eliminates the bias related to user-specific reference points. We can thus use it to study the properties of our normalized ratings in Eq. 5, and to provide an additional test for the relationship between quality, expertise, and stringency we have highlighted in Sections 4.2 and 4.3, this time without relying on the opinions of critics or award committees. Once again, we confirm that *i*) more expert users tend to be more stringent; and *ii*) high quality movies are rated by a more stringent audience.

Figure 3a compares the normalized stringency of users, s_j^{norm}, for those with a different total number of ratings on the platform, confirming our assumptions. Users with less than 150 ratings tend to be lenient and post higher ratings. Conversely, users who post more than 150 ratings are stricter. In particular, users with more than 300 ratings post, on average, ratings that are lower than average by a 0.1-star margin (on a 5 stars scale). This last group of users posts many more ratings than others and is thus over-represented in the ratings. In other words, a larger amount of movie ratings is posted by relatively more stringent users. As a direct result of the positive relationship between stringency and expertise, movies are more likely to be rated by an average population of stringent users.

Figure 3b shows the normalized movie stringency, s_i^{norm} for movies with different normalized ratings $\bar{r}_i^{norm}$. Movies with higher normalized ratings – and, thus, higher quality



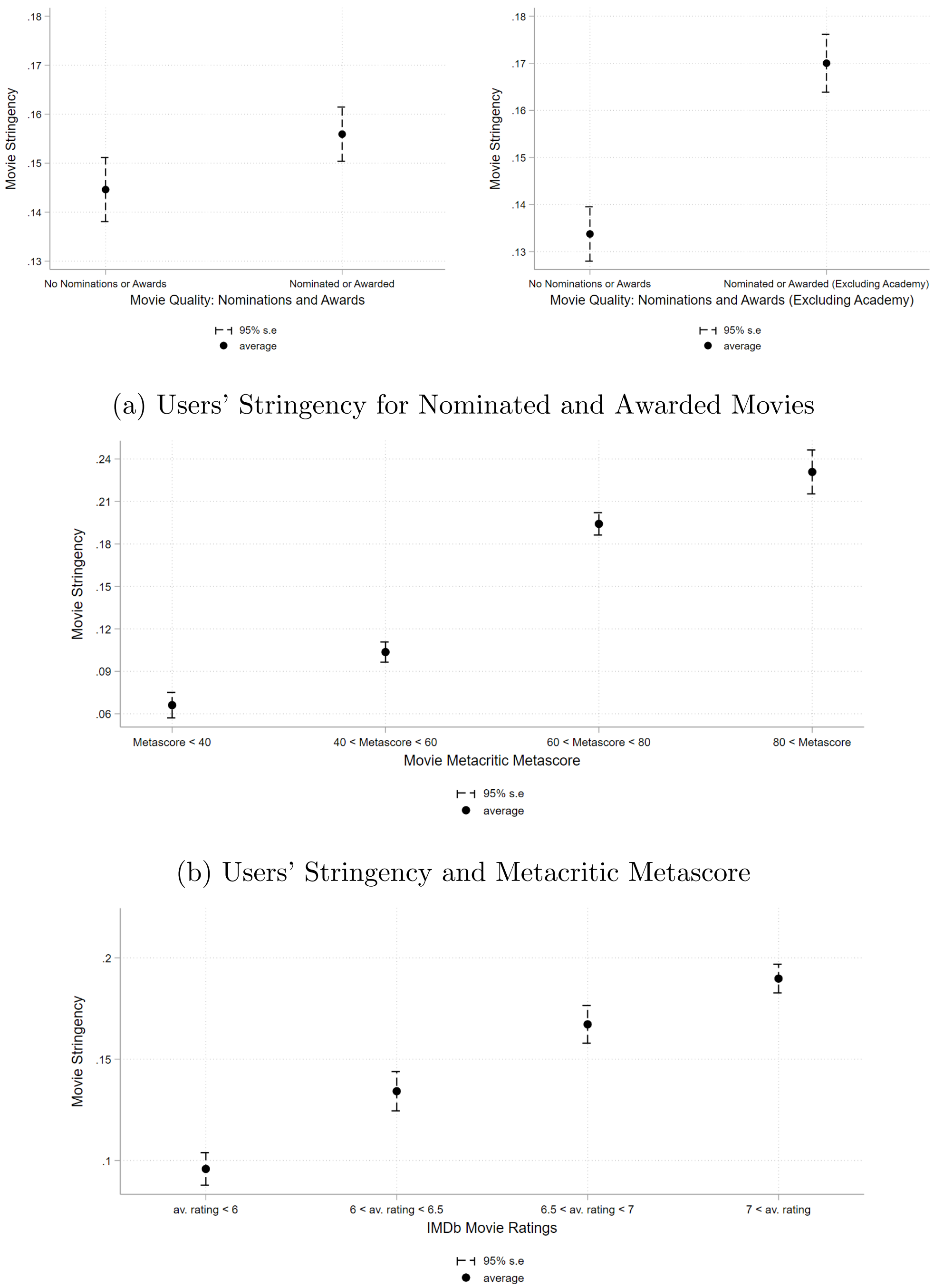
(a) Users' Stringency and Total Number of Ratings

(b) Users' Stringency and Average Normalized Ratings

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

– are rated by more stringent audiences. Accordingly, high quality movies are significantly penalized by the reference-dependent biases of the rating system. In terms of normalized MovieLens ratings, the highest quartile of movies (with $\bar{r}_i^{norm} > 3.5$) suffers almost a 0.15-star penalty compared to the lowest quartile of movies (with $\bar{r}_i^{norm} < 3$). This penalty equals one-third of the standard deviation of the normalized rating distribution, and is thus economically significant.

In Figure 4, we confirm the positive relationship between movie ratings and users' stringency using external measures of quality. Figure 4a shows that nominated and awarded movies tend to have a more stringent audience than movies with no nominations. Interestingly, in line with our empirical strategy, the effect is stronger when we focus on more obscure awards and festivals and exclude the Academy Awards, which are likely to attract a higher number of Non-Experts, in line with the mechanism explored by Rossi [2021]. Figure 4b focuses on movies with aggregated movie critics' reviews by Metacritic. Here we show



(a) Users' Stringency for Nominated and Awarded Movies

(b) Users' Stringency and Metacritic Metascore

(c) Users' Stringency and IMDb ratings

Figure 4: Exploiting User-specific Information to Debias Ratings: External Measures of Quality

that movies with better critics reviews tend to have a more stringent audience. In particular, movies with very positive critic reviews (with a Metacritic Metascore larger than 80) suffer almost a 0.20-star penalty compared to movies with poor reviews (with a Metascore lower than 40). For the sake of completeness, Figure 4c confirms the positive relationship between movie ratings and users' stringency with IMDb ratings. Of course, these ratings likely contain a reference-dependent bias. Yet, they provide a measure of quality that does not directly relate to the debiasing approach we used. All these pieces of evidence confirm that high quality movies are significantly penalized by standard rating systems that do not account for the users' expertise and different reference points. With our approach, we can adjust this bias and assign a premium to movies that are watched and rated by a stringent set of users.

5 Implications and Conclusion

In this paper, we investigate the consequences of consumer 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 endogenously expect more, and thus post more stringent evaluations. As a result, ratings compress qualities: high quality products' ratings are relatively too low – and, sometimes, lower in absolute terms – compared to those of their inferior alternatives.

We test our claims by using data from IMDb and MovieLens, two of the largest movie rating platforms in the world, and find striking support for them. Experts rate higher quality movies, as proxied by both nominations and awards to international festival and industry awards and movie critics reviews. Moreover, Experts 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. That is, by relative overestimating the quality of the lowest quality products, not only do ratings fail to hold all products to the same standard, but they do so in a particularly insidious way.

Next, we show that, however pervasive and problematic this bias is, correcting for it is rather straightforward. We algorithmically debias the ratings by exploiting the full history of users' ratings and mechanically equating users' stringency levels. Our correction does not rely on identifying Experts (for instance, to overweight their opinions), or to have a prior of which movies are actually 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 any thing, the movie market is fairly "egalitarian": certain – arguably high quality – movies (*e.g. "The Shawshank Redemption"*, *"The Lord of the Rings"* or *"Schindler's List"*) are 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 even 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.[15] With restaurants, and many other categories, we suspect consumer ratings likely reflect an "apple-with-oranges" comparison, even more so than we show they do with movies.

Industry players 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.[16] 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.

In a recent viral video[17], Freddie Wong, a filmmaker and food influencer, explains how he only eats at Chinese restaurants with a Yelp score of 3.5 out of 5. His reasoning echoes the one in this paper: by attracting mostly Chinese patrons (which here play a role akin to that of Experts in our paper), these restaurants are held to a much higher standard than some of their inferior (and, often, higher rated) competitors, whose patrons are not as knowledgeable. This is in line with Proposition 1. While clearly tongue-in-cheek, Wong's sentiment resonated widely, and Wong's original post has been reshared tens of thousands of times. Similarly, some Airbnb hosts have recognized that certain segments of consumers are more likely to leave negative reviews than others, irrespective of quality.[18]

We believe our results are important for managers, platforms and consumers alike. For managers, we suggest another dimension of potential segmentation that is 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

[15] The main empirical disadvantage presented by restaurant ratings, and the reason why we have not focused on them in our empirical analysis, 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 our quality-based self-selection theory: even the same consumer could rate in a way that negatively depends on prices. See Luca and Reshef [2021] for an empirical study of the impact of prices on restaurant ratings.

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

[17] See `https://twitter.com/fwong/status/1569736492247044103`.

[18] See `https://community.withairbnb.com/t5/Hosting/A-question-for-French-hosts-and-guests/td-p/891063`.

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 – for instance, Amazon, IMDb, Yelp and many other platforms all do.[19] As this elite group of reviewers rates on a different (and harsher) scale from everyone else, emphasizing their ratings without normalizing their scale (as we describe in Section 4.4) likely exacerbates the "apples-with-oranges" bias we highlight.

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, picky reviewers, which makes it look worse than it actually is? Or is it the other way around?"*). Of course, such inference is made complex by the fact that it requires consumers to have some information about the product, which they likely do not have a lot of if they are choosing to read reviews in the first place. At the very least, when consulting individual reviews is possible (as it is on most online platforms), consumers should internalize the fact that a lukewarm, or even negative, review from a "superstar" reviewer with thousands of ratings on the platform might not be as bad a sign as it looks (and the opposite for an enthusiastic review from a novice).

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 of most consumers, and colleges' grading policies are both transparent and widely debated.[20] 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.

[19] For example, see the Amazon Top Reviewers, or the Yelp Elite Squad initiatives: `https://www.yelp.ca/elite`.

[20] 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 peer institutions, such as Harvard or Columbia.

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Fast and slow learning from reviews. Technical report, 2022.

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

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.

Benjamin Bushong and Tristan Gagnon-Bartsch. Learning with misattribution of reference dependence. 2016.

Luis Cabral. Reputation on the internet. *The Oxford Handbook of the Digital Economy*, pages 343–354, 2012.

Christoph Carnehl, Maximilian Schaefer, André Stenzel, and Kevin Ducbao Tran. Value for money and selection: How pricing affects airbnb ratings. Technical report, working paper, 2021a.

Christoph Carnehl, André Stenzel, and Peter Schmidt. Pricing for the stars. *Available at SSRN 3305217*, 2021b.

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.

Kevin Chung, Keehyung Kim, and Noah Lim. Social structures and reputation in expert review systems. *Management Science*, 66(7):3249–3276, 2020.

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.

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

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.

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.

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.

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

Michelangelo Rossi. Quality disclosures and disappointment: Evidence from the academy awards. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 790–791, 2021.

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.

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

APPENDIX

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

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

as desired.

Proof of Corollary 1

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 2 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 **??**. By continuity in ψ, the expression remains positive whenever $\psi < \psi^*$, for some $\psi^* \in (0,1)$, completing the proof.

■

Proof of Corollary 2

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

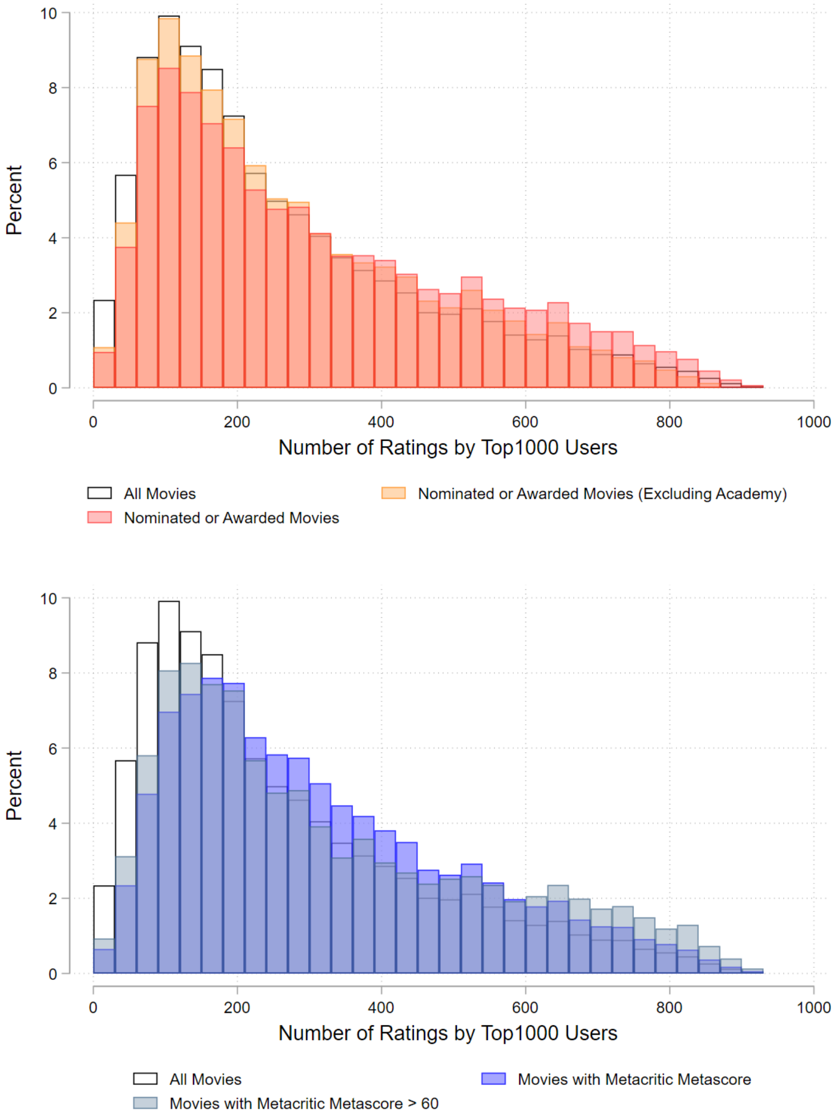
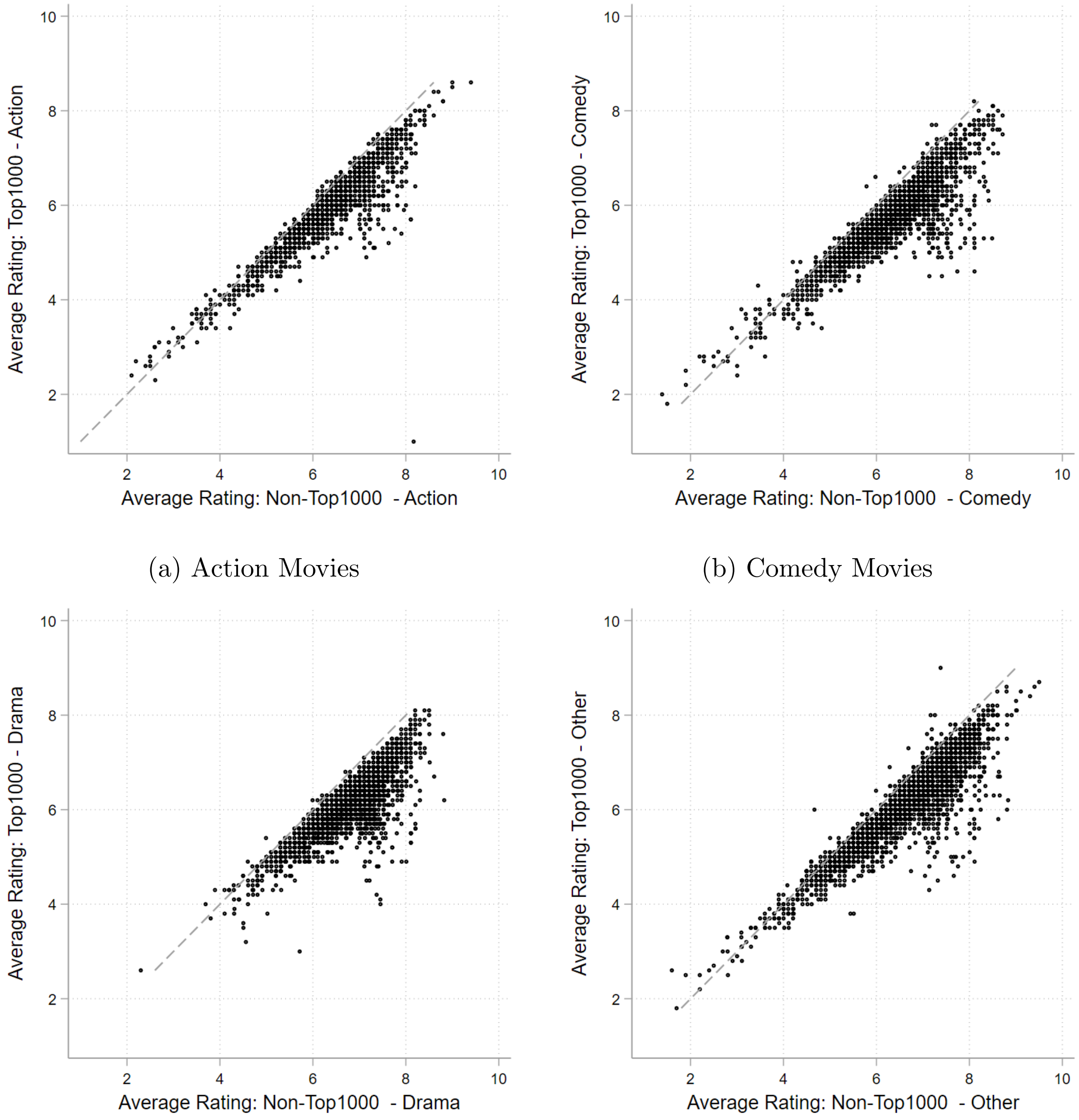


Figure 5: Distribution of the Number of Ratings Posted by Top1000 Users for Different Categories of Movies.



(a) Action Movies

(b) Comedy Movies

(c) Drama Movies

(d) Movies with Other Genres

Figure 6: Average Ratings for Top1000 and Non-Top1000 Users for Different Genres.

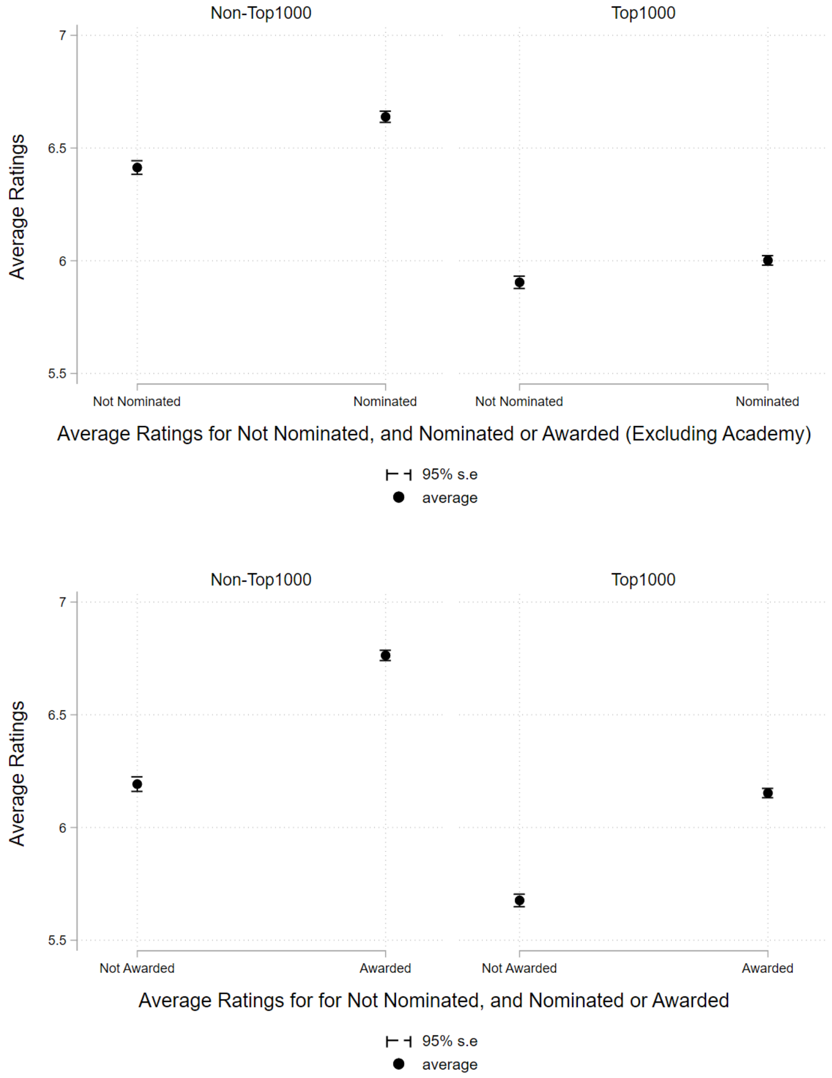


Figure 7: Top1000 Users Post Lower Ratings to Not Nominated, and Nominated or Awarded Movies