The Good, The Bad and The Picky: Consumer Heterogeneity and the Reversal of Movie Ratings

Tommaso Bondi and Ryan Stevens*
New York University

June 23, 2019

Please click HERE for the most recent version.

Abstract

We explore the consequences of consumer heterogeneity on online word of mouth. Consumers differ in their experience, which has two effects. First, experience is instrumental to choice: experts purchase and review better products than non-experts. Second, because of their superior choices, experts endogenously form higher expectations, and thus post more stringent ratings given quality. Combined, these two forces imply that the better the product, the higher the standard it is held to, the more stringent its rating. Thus, relative ratings are biased: low quality products enjoy unfairly high ratings compared to their superior alternatives. When this bias gets large, reputation needs not be increasing in quality. The bias needs not disappear, and can worsen, over time: products with unfairly high ratings mostly attract unexperienced consumers, reinforcing their advantage. We test our theory by scraping data from a well known movie ratings website. We find strong evidence for both of our hypotheses, and that this bias is quantitatively important. We then debias the ratings, and find that the new ones better correlate with the opinions of external critics.

*Email: tbondi@stern.nyu.edu and rls542@nyu.edu. We thank Hunt Allcott, Andrew Ching, Daniel Csába, Walker Hanlon, Brett Hollenbeck, Masakazu Ishihara, Michael Luca, Petra Moser, Franz Ostrizek, Venkatesh Shankar, Lena Song, Daniel Stackman, Shervin Tehrani, Larry White and particularly Luís Cabral, John Horton, Alessandro Lizzeri and Raluca Ursu for their helpful comments and suggestions. We also thank seminar participants at the Frank M. Bass FORMS Conference at UT Dallas, the London Business School Trans Atlantic Conference (Marketing), the NYU Stern Friday Workshop and the NYU Applied Micro Lunch. All errors are our own.

1 Introduction

In this paper we formalise, and provide empirical evidence for, the following idea: online rating systems reward lower quality products over their superior alternatives. This bias arises from the combination of two forces. First, more experienced consumers buy better products. Second, the utility consumers get from the products they purchase depends negatively on the average level of quality they are used to. That is, someone who often consumes very high quality products will rate a very good product as subpar, while the opposite is true for someone who is not as knowledgeable and picky.

The combination of these two forces imply that the reputation of high quality products - often purchased and reviewed by consumers who are used to high quality - will suffer compared to that of lower quality products whose consumers are more easily impressed and thus more lenient in their ratings. When this selection bias gets severe, aggregate ratings need not be increasing in quality, even in purely vertical markets, in which everyone would rank all products equally.

The following simple example helps fixing intuition: 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 finding the superior option. Non-experts, instead, pick A and B randomly, and therefore only 50% of them watch A. 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.[1] 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. Future consumers learning (naïvely) from these numbers might end up choosing B over A, even though they would have enjoyed A more. This simple fact has a key impact for long-term dynamics: if these consumers are more lenient in their ratings, then products – like B in our example – whose reviews were "unfairly" high initially will maintain this advantage, or even increase it, in the long-term. In other words, a rating bubble will form

[1] These features resemble our data in close detail: not only do connoisseurs review, on average, better products (where "better" will be formalised in the following sections), but their reviews are on average lower for every product, and further apart for different quality levels (in this example, $7 - 6 > 8.5 - 8$).

and the rating of B will keep increasing relative to that of A, so that the relative qualities of A and B will never be learned.

One obvious criticism is that if these biases in ratings are predictable, shouldn't consumers correct for them and draw proper inference about qualities? Indeed, in many real life scenarios we are aware of our recommenders' characteristics and can thus make inference about the informational content of their advice: if they developed a reputation for being easily satisfied, we will discount part of their enthusiasm. While this may be true for very educated consumers in physical markets, online learning makes this joint inference much more complicated. In most cases, consumers ignore the identities of the reviewers, and hence have limited capacity to extract the relevant information from their posted opinions. Moreover, these selection effects depend on quality itself, which is unknown in the first place. Therefore, while each rating is the combination of both the product's and the reviewer's characteristics, the latter are often ignored, producing an incorrect estimate of the former.

In this paper we provide empirical evidence for these phenomena, and complement it with a simple model which parsimoniously captures the main facts. In particular, we use extensive movie reviews data scraped from a well known website to show that *i*) consumers of different experience make different choices, *ii*) they also rate according to systematically different scales, and *iii*) this creates a sizeable bias in the ratings, and hence rankings, displayed by the website. However, from the econometrician's point of view, this bias is fairly easy to eliminate: if one category of consumers is more lenient in its ratings than another one, then we can just normalise all categories' ratings (i.e. equate their leniency), and compute new, corrected ratings and rankings. We perform this exercise and show that upon doing so, a measure of external validity of the website's reviews goes up: our new scores better correlate with a number of proxies of quality, for instance Academy Awards nominations.

Two unique features of our dataset play a key role. First, movies are uniformly priced (see for instance Orbach and Einav [2007]). This is important, as often prices, in conjunction with reviews, are used a signal of quality. Even if a fast-food restaurant has higher reviews than a Michelin starred restaurant, nobody would think the fast-food is better *in absolute terms*. In particular, one concern is that more discerning consumers are also willing to pay higher prices, and thus what we are capturing is simply the negative effect of price on reputation.

Second, the data comes in ideal form for our purposes: for each movie, on top of the overall score (the one saliently displayed on the website), an advanced search reveals more detailed averages concerning only specific subgroups of consumers; for instance, the movie might have an overall score of 8, while being rated on average 7.8 from under 30 viewers, 8.2 from women, and so forth. Of particular interest for us is the Top 1000 Users category:

in the rest of the paper, we will refer to them as experts and contrast their behaviour with that of average consumers. Moreover, the website specifies the number of consumers in each category reviewing each movie. Combined, these figures allow us to answer *i*) whether experience is associated with different (and, we will argue, "better") choices, *ii*) to more stringent ratings, and *iii*) whether the combination of *i*) and *ii*) creates a substantial bias in ratings, and *iv*) show what happens to ratings (and hence rankings) when we remove this bias.

We will model this adverse selection problem, together with its implications, in a vertically differentiated market. This requires justification, especially since our data deals with movie ratings – that is, an horizontal market. First and foremost, while understanding *why* experienced consumers might be more stringent in their ratings – we find the idea that that's due to their superior stringency very intuitive, and provide evidence for it, but also consider alternative explanations[2] – our conclusions, as well as the severity of the bias we identify, do not depend on the causes of this stringency, but only on its magnitude.

Second, we show in our data that, even controlling for horizontal characteristics (that is, analysing ratings only for a specific movie genre, or era), our results do not change. Third, and importantly, horizontal differentiation is often presented as the natural explanation for the mismatch between critics' and consumers' opinions: here, we show the existence of a completely orthogonal channel generating the same results.[3]

Despite its simplicity, this framework allows us to make some non-trivial and perhaps surprising predictions. First, ratings are highly context-dependent. This is due to the fact that products' ratings depend on their consumers' type, and this in turn depends on the goods' relative quality. In particular, improving the quality of the best of two goods needs not increase its ratings. Moreover, this will *favour* the absolute reputation of its worse alternative: a larger number of discerning types will realise it is the inferior option, and fewer of them will buy it. Even more perversely, this might favour the *relative* reputation of the inferior product. Context and framing effects on choice and utility have been documented extensively in individual decision making. Here, we generate these effects in aggregate while assuming them away at the individual level.

Moreover, popularity can be self-reinforcing: if we make the natural assumption that people with a lower understanding of the market rely more on socially generated information,

[2]For instance, is it because they have higher standards? Or because they are more able to detect failures when they occur? Or, do they experiment more than other consumers, hence occasionally end up experiencing products with which they match poorly?

[3]Winer and Fader [2016], in a reply to De Langhe et al. [2015], argue that the mismatch between critics' and consumers' opinions should not be seen as a problem, as weakly correlated signals of quality add more information to each other. This is, however, only valid if both signals are unbiased. If, as in this paper, low correlation is driven by the fact that aggregated consumer opinions are systematically biased, and consumers are unaware of this bias, then the conclusion does not hold.

then high ratings can bring an increasing share of them in the future, which would result in even more inflated ratings. If the best option prevails once, it will *a fortiori* prevail forever; however, in some cases the ranking of popularity will never reflect that of qualities. That is, we can not rely on long-term dynamics to correct initial biases. Indeed, our dataset includes movies that have received millions of reviews over more than a decade, and we still detect sizeable bias.

The rest of the paper is structured as follows: Section 2 describes the existing literature on the topic and its relation to our paper; Section 3 presents the model; Section 4 describes the data and Section 5 the empirical analysis, together with some important counterfactuals; we conclude and discuss future avenues for research in Section 6.

2 Related Literature

We relate to, and borrow from, different literatures. From a theoretical point of view, we combine well known ideas from the reference dependence (e.g. Kőszegi and Rabin [2006]) and the social learning (Banerjee [1992], Bikhchandani et al. [1992]) literatures. One interesting and related example of reference dependence is studied in Bushong and Gagnon-Bartsch [2016]: the authors consider a theoretical model in which as agents' reference points increase, their beliefs about quality become lower. For example, someone who has been flying business class for the last two decades might be convinced that the quality of the service is getting lower – while in reality he is simply becoming used to this standard of service, decreasing his positive surprise over time. The authors apply their model to study individual, but not social, learning.

We depart from the classical social learning literature in two ways: first, here consumers learn from ratings, not actions. That is, they not only see what previous consumers have chosen, but also a signal of their satisfaction. This is a fundamental difference, since the inefficiency highlighted in classic herding models would be solved if consumers could also rate their experience after taking action. Despite the superior communication technology, two elements in our model prevent efficient learning: consumer are both heterogeneous in their utility and choice, and naïve in their learning (a point already made in, among others, De Langhe et al. [2015] and Ursu [2017]). Heterogeneity skews the ratings; naïvete prevents the correction of this bias.

We also more specifically relate to the quickly growing body of research on online reviews. This research has focused, among other things, on quantifying the causal impact of ratings and rankings on choice (Luca [2016] and Ursu [2017]), systematic biases in ratings – both strategic (e.g. Chevalier et al. [2014], Nosko and Tadelis [2015] and Luca and Zervas [2016]) and non strategic (Li and Hitt [2008], Brandes et al. [2013], Acemoglu et al. [2017], , Ifrach

et al. [2014] and Besbes and Scarsini [2018]) – their complementarities with other forms of information (De Langhe et al. [2015]), and how to design more efficient platforms for social learning (Lafky [2014], Kremer et al. [2014], Che and Horner [2015]).

Closely related to the "ratings design" section of this paper is the work of Dai et al. [2012]. The authors argue that displaying only the mean of consumers reviews leave out a considerable deal of information (such as variance, median, trends) that is useful for consumers' decisions. Moreover, they advocate for a more sophisticated aggregation rule for individual opinions. Despite the completely different domain (restaurants vs movies) they find that most experienced reviewers have stricter standards, consistent with our model and data. However, they do not focus on the role played by experience in jointly shaping both choices and ratings, the main focus of our paper.

3 The Model

We now present a simple model to organise our main assumptions and results. Two important features should be kept in mind. The model simply assumes heterogeneous choices and ratings, but can be easily extended to one in which both arise endogenously. Moreover, our model deals with a vertical market. A model in which experts' stringency derives from horizontal forces would require substantially different assumptions. However, the quantitative importance of the bias we identify would remain unchanged.

There is a continuum of vertically differentiated products, whose qualities are distributed according to a continuous and smooth $G(q)$, $q \in [0,1]$. There is a continuum of consumers divided in two types, experts and non experts, totalling mass 1. Denote by $\psi \in (0,1)$ the proportion of experts. Both types choose exactly one product - that is, the outside option is 0 and hence never chosen. Experts choose according to $F_E(q)$, with density $f_E(q)$, non experts to $F_{NE}(q)$ ($f_{NE}(q)$).

We will assume that the two types self-select into buying different products. Specifically, we posit that experts' knowledge is instrumental in allowing them to buy superior products.

Assumption 1 (Self-Selection)**.** *For every $q \in [0,1]$, we have $F_E(q) \leq F_{NE}(q)$. Moreover, the two choice densities two satisfy the MLRP property:*

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

This assumption guarantees that experts are comparatively more represented for high quality products. Experts could, for instance, observe more precise signals of quality, or

have lower search costs.[4] We identify strong self-selection in our empirical section.

For now, as explained in the introduction, we abstract from other biases and focus on the case in which i) everyone reviews everything he experiences (that is, we only have selection on choice, not on rating conditional on choice) and ii) everyone reviews truthfully, that is, by stating his utility:

$$R_i(q) = u_i(q) \in [0,1], \quad \forall q \in [0,1], \ i = E, NE.$$

Our second key assumption is that experts are less satisfied for any given level of quality or, in other words, are more stringent.

Assumption 2 (Heterogeneous Stringency)**.** *For every $q \in [0,1]$, we have that*

$$u_E(q) \leq u_{NE}(q), \quad \exists \ (\underline{q}, \bar{q}) \subseteq [0,1] \ s.t. \ u_E(q) < u_{NE}(q) \ \forall q \in (\underline{q}, \bar{q}).$$

Just like in the case of choice heterogeneity, heterogeneous stringency is presented as an assumption here, but again, it is immediate to see that any model of reference dependence would yield this conclusion, as long as experts have higher reference points.[5] Our data offers striking support for this assumption, as we show in detail in Section 5.

For simplicity, we assume that the average rating displayed, $R(q)$, is linear in the individual opinions.

Assumption 3 (Linear Aggregation)**.** *The rating displayed by the platform is the average individual opinion, that is*

$$R(q) = \frac{\psi f_E(q) R_E(q) + (1-\psi) f_{NE}(q) R_{NE}(q)}{\psi f_E(q) + (1-\psi) f_{NE}(q)}.$$

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

We assume that $R(q)$ is the sole driver of social learning, and we study its static and dynamic properties. Since quality is in $[0,1]$, each individual utility lies in this space, thus

[4] For example, in the case of Probit models, this would require the natural assumption that experts' noise parameters be lower than non-experts'.

[5] A theoretically distinct, but plausible, explanation would be that this vertical market is a cross-section of within a set of horizontal markets. Experts experiment across horizontal markets, and hence are more likely to end up with some bad matches. According to this last interpretation, not every expert's rating would be more stringent, but the average of their ratings would, since it would include the ratings of some experts whose taste does not match the products purchased.

$R(q)$ also lies in this space. We will assume an outside option of 0, or equivalently, that each future consumer wants to buy (exactly) one product. Because of this assumption, we will mostly be interested in $R'(q)$; in other words, for every two levels of quality q_1 and q_2, we are interested in the difference in their ratings, $R(q_1) - R(q_2)$. Let's start with a couple lemmas that will simplify notation in what follows.

Lemma 1. *We can assume $R_{NE}(q) = 1$, that is, non experts do not recognise quality, without loss of generality.*

Proof. The only thing that matters for choice are differences in ratings. Consider a generic couple of functions $(R_{NE}(q),\ R_E(q))$. It is immediate to show that the new couple given by $(1,\ R_E(q) - R_{NE}(q) + 1)$ leaves $R(q_1) - R(q_2)$ unchanged, $\forall (q_1, q_2)$. ■

Lemma 2. *We can assume $f_{NE}(q) = 1$, that is, non experts choose randomly, 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 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

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

where $\psi' := \frac{\psi\alpha}{\psi\alpha+(1-\psi)}$. Note that $\psi' \in [0,1]$ by construction. 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 the admissible parameters. This completes the proof. ■

Given these two assumptions, we can work with

$$R(q) = \omega(q)R_E(q) + (1 - \omega(q)), \quad \omega(q) := \frac{\psi f_E(q)}{\psi f_E(q) + (1 - \psi)}, \tag{1}$$

where $\omega'(q) > 0$. Moreover, we assume that $\omega''(q) < 0$: this is consistent with models of discrete choice, including the Logit and the Probit, and with theories of information acquisition in which information has a convex cost.

Therefore,

$$R'(q) = \underbrace{\omega'(q)}_{\text{change in \% of experts}} \cdot \underbrace{\big(R_E(q) - 1\big)}_{\text{loss per experts}} + \underbrace{\omega(q)R'_E(q)}_{\text{increase in experts' utility}}. \tag{2}$$

The equation admits nice intuition, and formalises the main point of our paper. Marginally increasing quality has two opposite effects. On one hand, each of the non experts remains equally satisfied, while the $\omega(q)$ experts become more satisfied by $u'_E(q) = R'_E(q)$. On the other hand, their proportion increases by $\omega'(q)$, and this multiplies an individual loss of $R_E(q) - R_{NE}(q) = R_E(q) - 1$. We are interested in the monotonicity properties of $R(q)$, that is in necessary and sufficient conditions for $R'(q) > 0$.

To this end, note that $R'(q) \geq 0$ if and only if

$$R'_E(q) \geq \frac{\omega'(q)}{\omega(q)}\big(1 - R_E(q)\big), \tag{3}$$

that is the increased satisfaction per experts is greater than their percentage increase multiplied by the individual self-selection losses.

Assume that $R_E(q) = q$[6]. This assumption simplifies the notation in what follows. However, there is nothing special with linear utility in this context: any increasing function of q yields the same qualitative predictions.

Equation (3) becomes

$$1 \geq \underbrace{\frac{\omega'(q)}{\omega(q)}\big(1 - q\big)}_{\text{Self-selection loss}} =: L(q). \tag{4}$$

We can then prove our first result.

Theorem 1. *With the aforementioned assumptions, reputation is increasing in quality if and only if* $\frac{w'(0)}{w(0)} < 1$*, and U-shaped otherwise. That is, in this latter case, there exists a* q^* *such that reputation decreases in* $[0, q^*)$ *and then increases in* $(q^*, 1]$*.*

[6]Here, this is merely a technical assumption. In particular, we do not imply that experts ratings are perfectly accurate, while non-experts' are not.

Proof. We want to characterise the q's such that $L(q) < 1$. First notice that $\omega''(q) < 0$ is sufficient for $L(q)$ to be decreasing:

$$\frac{\partial L(q)}{\partial q} = \frac{\omega''(q)\omega(q) - \omega'^2(q)}{w^2} \cdot (1-q) - \frac{\omega'(q)}{\omega(q)} < 0.$$

This is intuitive: the loss per experts, $1-q$, is decreasing, and the marginal increases in the proportion of experts is also decreasing, by concavity of $\omega(q)$.

Moreover, we have that

$$\lim_{q\to 0} L(q) = \frac{\omega'(0)}{\omega(0)}, \quad \lim_{q\to 1} L(q) = \lim_{q\to 1} \frac{\omega'(q)}{\omega(q)} \cdot (1-q) \to 0.$$

Therefore, $L(q) < 1 \ \forall q \in [0,1]$ if and only if $\frac{\omega'(0)}{\omega(0)} < 1$. In this case, reputation is monotonically increasing in quality. If instead $\frac{\omega'(0)}{\omega(0)} > 1$, denote by q^* the unique quality such that $L(q^*) = 1$. Existence and uniqueness of such q^* are guaranteed by the continuity and monotonicity of $L(q)$, and by the fact that $(L(0)-1)\cdot(L(1)-1) < 0$. Then, reputation is decreasing in $[0, q^*)$ and increasing in $[q^*, 1)$. ■

In words, the theorem states that when self-selection is strong enough, products of intermediate quality (that is, whose quality lies in an interval including q^*) suffer an unfairly low reputation, relative to inferior alternatives: their quality is high enough to attract some experts, but too low to satisfy them. This ends up penalising them compared to inferior products that are not as purchased by experts.[7] The comparative statics properties of q^* shed additional light on this phenomenon. Loosely speaking, a low q^* (including the limit case in which reputation is increasing) reduces the severity of the problem. We are interested in how q^* changes as *i*) the proportion of experts increases, *ii*) buyers become more heterogenous in their choices and *iii*) buyers become more heterogenous in their ratings. The next Corollaries formalise these results.

Corollary 1. *The inflection point q^* is increasing in ψ if and only if the following condition is satisfied:*

$$\psi < \frac{1 - \sqrt{f(q^*(\psi))}}{1 - f(q^*(\psi))},$$

or, in other words, when ψ is small compared to q^*.

It is important to notice that one can interpret choice heterogeneity between more and less experienced consumers as (being proportional to) the degree of vertical differentiation

[7] Scholars have been pointed out that in recent years, evidence has accumulated for a *vanishing middle* in a variety of markets. This is consistent with our model. We are not, of course, claiming that the vanishing middle is solely determined by reputational dynamics, as there are likely more important market forces involved.

in the market. It is precisely this heterogeneity that lies behind the varying proportion of experts purchasing products of different qualities. Therefore, our bias is more severe precisely in markets in which incorrect social learning is more costly. Non-monotonicity of reputation in quality gives a particularly stark result, and relies on the specific assumptions made on the strength of both self-selection and rating heterogeneity. The data will show that this benchmark, though seemingly extreme, is not quantitatively unrealistic. Moreover, it is important to notice that a similar lesson from the model holds much more generally.

Theorem 2. *For every q_1, q_2 such that $q_1 < q_2$, we have that the $R(\cdot)$ mapping is a contraction:*

$$\left|R(q_2) - R(q_1)\right| < q_2 - q_1.$$

That is, products of higher quality are *always* penalised when compared to lower quality alternatives. This is because consumer self-selection mitigates the individual gains from quality. One could think pathological outcomes are prevented as long as $R(q_2) - R(q_1)$ is positive. However, in a variety of natural applications we might have that future agents choose q_2 only if it looks substantially better than q_1 (say, because of horizontal differentiation, or price), so that this contraction result has implications for choices made in future periods.[8]

A key thing to notice in this static framework is the potentially problematic interactions between this bias and the policy measures taken by several online platforms to increase the accuracy of their publicly generated ratings. For instance, both Amazon and Yelp weight the ratings of experienced reviewers more heavily than those of non experts. There are natural rationales for this: non experts' ratings might be less reliable, and part of them might just display fake information posted strategically by sellers. However, these motives are limited here, and in many similar contexts. First, fake reviews would not do much harm for products already being reviewed thousands, if not millions, of times; indeed, Chevalier et al. [2014] show that sellers give up trying to influence their own ratings in these contexts. Second, although in our model experts are reliable and non experts are not, this unreliability cancels out when considering *relative* reputations, arguably the main object of interest for social learning[9]. Moreover, we highlight that these policy measures can backfire, as far as our

[8] It is important to notice that this type of mislearning of quality from the behaviour of predecessors is opposite to what highlighted in Gagnon-Bartsch and Rabin [2016]: there, the key assumption is that agents neglect that their predecessors acted upon observing those that came before them, and not solely relying on their private information. This leads agents to overestimate the precision of the predecessors' private information: as a result, they will think that herding in the previous period corresponds to one option being substantially better than the other, effectively overestimating quality differences. While their model explains the (excessive) presence of superstars, ours points out to the fact that some stars might not shine, due to the adverse selection of consumers they face. Another important difference is that in their model distortions arise dynamically, as a consequence of naïve learning. In ours, ratings are biased even in the static benchmark, although the welfare consequences of this fact clearly depend on the (mis)learning dynamics.

[9] This is especially true in situations in which consumers have a 0 outside option, so that they effectively learn through *relative* ratings, and not absolute ones.

self-selection bias is concerned.

Corollary 2. *Assume that platforms overweight each expert's review by a factor $\gamma > 1$. Then, q^* goes up whenever the condition in Corollary 1 is met.*

Proof. Overweighting experts by γ means that now

$$R(q) = \frac{\gamma\psi f_E(q)R_E(q) + (1-\psi)f_{NE}(q)R_{NE}(q)}{\gamma\psi f_E(q) + (1-\psi)f_{NE}(q)}.$$

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

$$\frac{\gamma\psi}{\gamma\psi + (1-\psi)} > \psi.$$

Applying Corollary 1 proves the result. ■

4 Data and Empirical Strategy

4.1 Data

The data consists of movie ratings submitted anonymously by individual users on a well known movie review website. Ratings must be a whole number between 1 and 10. Unlike on other online platforms, such as Yelp, users are not able to observe the individual ratings of other reviewers on the site. Instead, they can only see summary statistics of aggregated ratings.[10]

The website has lists for the Top 250 ranked films on the website, as well as an overall Top 250 ranked list and a set of Top 250 lists separated by genre. Five different lists were scraped, a list of the Top 250 ranked movies overall and a set of lists of the Top 250 ranked movies by genre for the following genres: Horror, Comedy, Drama and Action. We were worried that "popular" products may bias our results, as these could be the places where non-expert and expert preferences diverged the most. To account for this, a sixth scrape run was done for a set of lists which ranked movies by their box office revenue by year. We scrapped the Top 100-1000 movies by box office revenue, by year for movies released between 2010 and 2017. Due to the limited speed of the algorithm, we choose a random sample of 50 movies for each year for our analysis.

Conveniently for us, a more advanced search (which is extremely unlikely to be performed by users) reveals a variety of scores for each movie. Not only do we observe the overall mean

[10]Thus, social conformity and / or social reputation leading to non-truthful reporting are not an issue with our data, which motivates the proxying of ratings with utilities in our model.

score – that is, the $R(q)$ in our model – but also the scores given by different categories of reviewers. For example, we observe the ratings given by women and men; by younger and older users; by US and non-US reviewers; and so forth. Key to our analysis, the website splits out ratings across two groups of reviewers: (1) Top 1000 and (2) Non-Top 1000 reviewers. Importantly, the set of Top 1000 reviewers is anonymous: that is, even users belonging to it are not aware of this fact. For this reason, we do not have to worry that differences in choices and ratings are motivated by reputational concerns, such as a desire to be perceived as tough graders, or possessing a sophisticated taste.

We will classify the Top 1000 reviewers as experts and the Non-Top 1000 reviewers as non-experts. Given that the number of experts on the website is arbitrarily capped at 1000, which is certainly an underestimate of the number of reviewers on the sites who have experience reviewing films, as a robustness check we will also use age as a proxy for expertise.[11]

Given that we are ultimately interested in how quality and rankings are correlated, we connected our movies to an external measure of quality, that is, whether a movie was nominated for an Academy Award. Proxying for quality using external awards is a standard procedure when dealing with consumers reviews, see for instance De Langhe et al. [2015] and citations therein.

In summary, the sets of variables associated with our scraping runs, in addition to the Academy Award information can be divided into three sets: (1) Reviewers Demographics, (2) Movie Characteristics and (3) Ratings. Broadly speaking, we will investigate how ratings are not solely determined by Movie Demographics, but rather depend on the set of reviewers for each movie, as predicted by our model. Table 1 gives a description of the variables used in the analysis.

Prior to the analysis, we used two cleaning steps. First, we dropped repeats of movies, since the same movie could appear across scraping runs. For example, a comedy movie in the Top 250 overall ranked list will also show up in the Top 250 comedy ranked list. Second, we were worried about foreign films that were not widely distributed, due to both cultural differences and promotional reviews. To address this, we dropped any movie in which the proportion of US reviewers was less than 10%. This was not a very strict exclusion, as it caused us to drop only $\sim 1\%$ of the total number of ratings in our dataset.

Table 2 gives summary statistics of our final dataset for the Reviewer Demographic variables, in addition to the sample size for each scrape. We have $1,790$ movies in our final dataset. The average number of reviews by movie is very high (for instance, it is approximately $441,000$ for the Top 250 overall movies). The demographic skews male and

[11] This choice is justified by the strong correlation in both choice and ratings between older and Top 1000 reviewers. We will show this correlation later on.

younger than 45. Furthermore, while there is a US bias on the site, on average only 29% of reviewers are from the US. Finally, outside of the number of reviews, there are no large differences in demographics across the scrapes.

4.2 Empirical Strategy

Our empirical strategy is as follows: first, we verify that the model's assumptions are valid. We proceed in two steps showing that (i) experts choose different (and, we will argue, better) products and (ii) conditional on choices, experts rate products more stringently. We refer to (i) as *choice heterogeneity* and (ii) as *ratings heterogeneity*.[12] We find strong evidence in the data for both of these hypotheses. Second, we study the implications of this self-selection bias: we begin by debiasing the existing ratings in different ways, and constructing a new ranking made of the corrected ratings. We study in detail the properties of this newly constructed ranking. How does it compare to the one currently displayed by the website? We find that our ranking better correlates with external awards, and Oscar nominated movies gain on average 16 positions. This is not tautological, since the conclusion remains true even when we assign little (but fixed) weight to experts' opinions. We also study two of the key implications of our model. First, we show that selection has the potential to generate non-monotonicity between reviews and quality. Second, we show that ratings are a contraction in quality, i.e. the differences in quality are greater than the differences in ratings displayed on the website.

Last, we begin investigating a broader question: is there a wisdom of crowds effects? In other words, how would ratings look if we were to completely neglect non-experts, and instead only consider the opinions of the Top-1000 reviewers? Is it desirable that everyone posts his ratings, or would ratings and rankings look more reasonable - as proxied by their similarity to external metrics - if we were to disregard some, arguably less informed opinions? Answering this question is important, but challenging. The reason for this is that it is not clear whether external ratings - mostly consisting of the aggregated opinions of professional critics - can be thought of as measure of objective quality without ignoring horizontal considerations: it might simply be that our top-1000 reviewers better correlated with Oscars purely due to this. However, some data analysis rules out this possibility: our debiased ratings look more

[12] Since we only observe reviews, we conflate choices to review with choice to watch a film. One possible criticism of our approach is we only measure the choice to review a film, and we are not measuring conditional on choosing to watch a film, how do experts and non-experts differ in their willingness to write a review? For instance, experts may review all films, whereas, non-experts only review the films that they like. This criticism does not undermine our results, but it changes their interpretation. Regardless of how the distributions of choices and ratings are generated, so long as different films are being reviewed in different proportions by experts and non-experts, the selection effect highlighted in our model biases rankings.

in line with Oscar nominations when the opinions of non-experts are taken into account.[13]

5 Results

5.1 Choice Heterogeneity

We first investigate differences in choices between experts and non-experts. To investigate choice heterogeneity, we test whether experts choose to review higher quality films than non-experts. One way to show this is to see if the proportion of experts reviewing a film increases if the film is nominated for an Academy Award. The empirical specification is as follows:

$$\begin{aligned}\text{prop\_Top\_1000}_i = \alpha + \sum_{g \in genre\_set} \theta_g \mathbb{I}(\text{movie\_genre}_i = g) + \theta_{year}\, \text{movie\_year}_i \\ + \theta_{us} prop\_US_i + \theta_n num\_reviews_i + \beta\, \text{oscar\_nom}_i + \epsilon_i\end{aligned}$$

Our covariates include α a constant, θ_g a set of genre fixed effects indexed by genre g, θ_{year} the effect of year the movie was released, θ_{us} the effect of proportion of US reviewers, and θ_n the effect of the total number of viewers. Our object of interest is β, which measures the effect of a movie being nominated for an Academy Award. Specifically, it measures the increased proportion of experts reviewing an Academy Award nominated movie relative to a non-Academy Award nominated movie. We assume errors are drawn i.i.d. across movies from a logistic distribution. Because the dependent variable is a proportion between 0 and 1, we estimate a fractional logistic regression model.

Table 3 shows the results of this regression, where the first and second model differ based on controlling for genre fixed effects. Controlling for our covariates, a nominated movie increases the proportion of experts by 0.09. Since the average proportion of experts reviewing a movie is $\sim$0.35, this is nearly a 25% increase in experts reviewing a film. This provides initial evidence that experts tend to choose higher quality movies.

Given that the website caps the number of experts at 1000, our regressions detailing the top 1000 experts choices and quality of a movie is subject to censoring biases. This is magnified by the fact we are observing a particular community of experts. If the site has drawn attention to certain films, the top 1000 users who are most likely very active users on the site might be more likely to rate a film. However, the larger population of experts

[13]This is comforting for two reasons: first, because as explained before it rules out the possibility that horizontal differentiation is the main driver of our results. Second, and more importantly, because it provides evidence that there is value in the common provision of reviews, even by less experienced consumers.

beyond the top 1000, may be less moved by these sorts of attentional biases. Thus, as a robustness check, we investigate the relationship between reviewers aged 45 years or older and nominated films. The website gives us count data on the number of individuals aged 45 years or older that review a film, thus we can get around our issue with the proportion being capped. We use this subpopulation as a robustness check because age is a plausible mechanism in the formation of expertise. Older people have the ability to build up more experience over time, and thus are more likely to have similar preferences as experts. We will show later on that this subpopulation of reviewers have similar preferences to the experts on the site.

For now, assuming this similarity in preferences across experts and older reviewers, do these older reviewers choose to review higher quality movies? We redo the same analysis as before, replacing the dependent variable with the total number of reviewers aged 45 or older.[14] As shown in Table 4, Academy Award movies increase the number of older reviewers significantly. [15] Moreover, the coefficients for our other variables make intuitive sense.

5.2 Rating Heterogeneity

So far we have established that experts tend to review higher quality movies. We now ask whether experts tend to be more stringent relative to non-experts in their ratings. Our data is ideal for this purpose: we can get at this question by simply comparing mean scores for experts and non-experts for each film in our sample. Figure 1 plots the mean score of Top 1000 reviewers relative to the mean score of Non-Top 1000 reviewers, where each point is a single movie. If experts and non-experts had similar stringency, the points should be centered around the 45 degree line. However, we see that this is overwhelmingly not the case: nearly all points are below this line, indicating that on average experts rate movies lower than non-experts. Sometimes this difference can get extreme: for some movies, the difference in experts and non experts ratings can be as large as $5-7$ points out of 10! Given the meaning of these numbers for social learning - with a movie rated 8 or above usually being a "must see", while a rating of 6 usually stopping users from watching a movie - these differences are substantial.

These plots tell us nothing about how higher ranked and lower ranked movies are reviewed. Remember, our first five scrapes were from lists of Top 250 ranked movies. The effect could be resulting from experts tending to prefer higher ranked versus lower ranked movies on our website. As shown in Figure 2, this is not the case. Experts are always

[14] Since the dependent variable is continuous and unbounded from above, we use OLS to estimate our parameters.

[15] We are explicitly controlling for movie year, so we are not purely capturing the fact that Academy Awards winners are older than the average movie.

reviewing movies lower across all rankings. As a robustness check, we also scrapped a Top 250 list for television series. Figure 3 shows that our effect is more general than movies, as experts are more stringent than non-experts across rankings. The same goes for rankings by genre, in which horizontal differentiation plays a much more limited role.

Another way to see this same result is to plot mean reviews by age bins. Figure 4 shows this plot. As stated earlier, reviewers aged 45 or older tend to choose similar movies as those top 1000 reviewers, or experts. However, notice that the these two groups: experts and older reviewers do not give exactly the same scores. Comparing the lines in Figures 2 and 4 are not the same, experts are even more stringent than the average aged 45 or older reviewer. Thus, while these people are a good proxy for experts, age is not the only story. This seems plausible as gaining expertise is not simply about mere exposure and time, but requires costly effort to experiment and learn.

One of the mechanisms proposed to explain this stringency effect was that non-experts have a higher ability to discriminate quality. For example, consider a consumer choosing which movie to watch during the weekend. Say movie A is chosen. Conditional on it being chosen, expectations over its quality must be fairly high. Then, they can be revised downwards as bad signals are observed: some deficiencies in acting performances, aesthetics heavily borrowing from other movies, and so forth. All that we are assuming is that the ability to detect such failures is increasing in expertise. One way of investigating this is comparing mean reviews for experts and non-experts depending on whether the movie was nominated for an Academy Award. Nominated movies should have fewer deficiencies, and hence ratings across categories should be closer for them. Consistent with this, Figure 5 shows that non-experts give on average a 0.5 point bump to nominated films, whereas experts give a 1.6 point bump to nominated films. Table 6 confirms our story about how quality correlates with these two groups reviews. The Academy Award coefficient is significant and positive for experts, but insignificant for non-experts. Importantly, note that experts are more stringent even for nominated movies, the ones they relatively like more; again, this points against horizontal preferences being the main driver of our results.

One possible story could be that experts are not reporting truthfully, but taking into account this stringency effect and "punishing" films by reporting ratings lower than their perceived quality. Figure 6 shows the distribution of reviews between groups. As can be seen, while experts do tend to give more 1's than non-experts, there is an overall shift in the distribution from giving 9's and 10's to lower values like 6's, 5's etc. This also confirms the idea that experts are simply reviewing on a different scale than non-experts. The fact that most of the action happens on 9's and 10's is intuitive: experts can separate very good movies (say, 8's) from absolute masterpieces, while non-experts do not discriminate as much for high quality movies.

Notice how so far we have focused on the fact that experts' and non experts' ratings differ in their means, not their variances. The latter might also be natural to expect, at least if we assume that experts receive a more precise signal about a movie's quality. If this were the case, we would expect the variance, by movie, of expert reviews to be lower than the variance of non-expert reviews. To investigate this, we compute the variances of expert and non-expert reviews within each movie, then take the average variance across all movies. The results do not point towards this signaling story. The average variance of experts' reviews is 6.04, whereas the variance of non-experts' reviews are 3.66. As a robustness check against outliers, we compute two alternative measures of dispersion. First, we compute the average difference in the 90th and 10th percentile for both experts and non-experts by movie. Second, we compute a pooled variance measure for experts and non-experts, where a specific movie's variance is given a higher weight the more reviews there are for the movie. Both of these alternative measures confirm our original finding. First, the average 90/10 percentile difference for experts is 6.26 compared with an average difference of 4.40 for non-experts. Second, the pooled variance of experts is 4.84 compared with a pooled variance for non-experts of 3.06.

To summarize our results thus far, experts tend to choose movies that are of higher quality. For virtually every film, experts are more stringent in their ratings, leaving considerably lower reviews than non-experts. These two facts combined confirm our theoretical predictions: higher quality products' reputations are downward biased compared to low quality products'. This adverse selection mechanism is quite pronounced: we will show that in our sample, reputation does not appear to be monotonically increasing in quality, in line with the more extreme result in our model.

5.3 Counterfactual Ratings

The natural next key question becomes: after identifying this bias, can we eliminate it? We propose a simple method to debias ratings, and then show the new rankings resulting from this correction. Furthermore, we show these new rankings are substantially different from the website rankings. Finally, we show our debiased rankings are more correlated with our quality measure than the website rankings. Note, we will not use any information about the movie to generate our new rankings. Instead, we simply rescale our ratings based on the type of reviewer ranking the product. This is consistent with the story we laid out at the beginning: often, *who* is reviewing is just as important as *what* is being reviewed. Our procedure attempts to control for the *who*, so to isolate the *what*.

Since the self-selection bias is the combination of heterogeneity in choices and ratings, one can eliminate it by shutting down either channel. To shut down the latter, one would have to compute the average difference in stringency between these two categories, and increase

experts' reviews by this amount. To shut down the former, which is the approach we follow, one simply has to fix weights given to experts and non-experts in determining the final rating across movies. That is, if two movies had different reputations simply because the first had a more experienced - and hence stringent - crowd than the latter, we eliminate this difference by essentially equalizing the proportion of experts and non-experts in the two counterfactual ratings. We do so by fixing different weights for experts. Clearly, this leaves us with a degree of freedom: do we want every movie's rating to look as if 70% of its reviewers were experts? Or is 30% better? We will leave this parameter free and compute new rankings as a function of it. This is also key for investigating wisdom of crowds effects: does giving a very high weight to experts, hence essentially disregarding common consumers' opinions (which effectively represent a large portion of reviews, in the majority of markets), lead to a higher correlation with external measures of quality?

Formally, let $newRating_i$ be the new rating for movie i, and $\gamma_E \in [0,1]$ be the weight given to experts[16]:

$$newRating_i = \gamma_E * \text{Top\_1000\_mean} + (1 - \gamma_E) * \text{nonTop\_1000\_mean}$$

Thus, if $\gamma_E = 1$ only the mean score of experts reviews are used, whereas if $\gamma_E = 0$ only the mean score of non-experts reviews are used. We then take the analysis one step ahead and see what happens when performing comparative statics on this parameter. After computing a new rating for each movie, we then re-rank the movies. We do this analysis on the Top 250 Overall list.

The key issue is whether our bias-corrected reviews give higher rankings to higher quality movies, as proxied by external indicators of quality. To do so, we re-run our original regression with the dependent variable being the difference between rankings in the original Top 250 list and our new debiased ones.[17] Table 7 shows our results: Academy Award nominated movies are 16 positions higher in our debiased ranking than the original ranking.

We believe this is an important finding. The widespread disagreement between online reviews and proxies for quality, such as consumer reports and awards, has long been a theme of discussion. One natural explanation for such misalignment are taste differences. In our theoretical model, we show that this is only half of the story, and that the erroneous aggregation of consumers' opinions can complement taste differences in causing this discrepancy. Our data confirms this: we have increased the alignment between these two vectors of opinions without recurring to horizontal considerations. As an extreme case, take correlations

[16] In other words, our counterfactual exercise corresponds to substituting each movie specific $\gamma_{i,E}$ with a γ_E that is fixed across movies.

[17] No genre controls were used because of the reduction in our sample size.

across two new counterfactuals, $\gamma_E \in \{0, 1\}$. As previously stated, these are the cases when only non-experts or experts are used in creating the counterfactual ranking. Table 8 compares the correlation across all our rankings. The correlation between the Top 1000 rankings and the website ranking is much lower than the Non-Top 1000 rankings. Furthermore, our rankings giving equal weight to both groups substantially advantages the experts, as it is highly correlated with their rankings but not the non-experts. Again, this is comforting: if unbiased ratings better reflect quality, and experts are able to better detect quality than non-experts, then the correlation patterns should be exactly the ones we find.

Our model says more than simply that bias corrected reviews improve quality, there are two other testable implications from our model. First, reviews may be non-monotonic in quality. Second, reviews are less dispersed than quality, i.e. reviews are a contraction relative to underlying quality.

We first show that this non-monotonicity result holds for certain genres of films. Figure 7 plots the relationship between our original mean scores and our bias corrected mean scores, for each of the six scrapes that we run.[18] As can be seen for Top 250, Horror and Comedy, we have a non-monotonic relationship. While the result is not true for all genres, we see that monotonicity between ratings and quality is not a foregone conclusion of review systems.

With regards to our contraction result, a bit of discussion is in order. When we say "reviews are a contraction relative to quality", we mean the relative difference in qualities between two films is larger than the relative mean review scores. We have already shown that our bias corrected review scores better reflect underlying quality. Thus, a simple test of the contraction result is to take each pair of movies, take the ratio of the difference in the bias corrected review score relative to the original review score. If this quantity is greater than 1, then this is a contraction. To test our contraction result, we construct a "contraction ratio". For each of our six scrapes seperately, we take each pair of movies, calculate the ratio described above, and then take the mean ratio within each genre. Figure 8 shows the average "contraction ratio" across each scrape.

5.3.1 Wisdom of Crowds

Giving experts equal weights tends to debias reviews and give a significant bump to higher quality films. However, as we have argued before, this may be mechanical as experts may just be a similar population as Academy Award critics. If this is the case, this similarity should be increasing in γ_E. To this end, we repeat our rankings improvement regressions

[18] We construct these plots by first calculating our bias corrected mean scores by setting $\gamma_E = 0.5$. We then normalize mean scores (original and bias corrected seperately) within each category. Following that, we regress the normalized original mean review score on the normalized bias corrected mean score and the normalized bias corrected mean score squared. Each point represents a movie. The shaded regions are bootstrapped 95% confidence intervals.

from earlier with $\gamma_E \in \{.10,\ .20, \ldots, .90, 1\}$. That is we fix a γ_E, compute a debiased ranking, then take the difference between the original and debiased ranking, and finally run our regression shown in Table 7. Figure 9 plots the coefficients (oscar_nom) for each of these regressions, i.e. for different values of γ_E. Thus, as you move from left to right in the graph, experts reviews are receiving higher weight. Interestingly, and key for the interpretation of our results, we find that giving higher weight to experts matters very early on, but the marginal value of adding more experts goes to zero.

This story is consistent with our findings: assume rankings are debiased, so that average ratings are constant across groups. Experts are more likely to identify quality, i.e. the aggregate distribution of expert reviews are more precise, but their numbers are capped at 1000. On the other hand, non-experts are less likely to identify quality individually, i.e. the aggregate distribution of non-expert reviews are less precise, but there are many thousands of non-experts. Are they so imprecise that using their opinions, and not only the experts' ones, make the overall, correctly aggregated rating less precise? The answer seems to be negative in our dataset. This is not surprising, since *i*) their number is so large that individual imprecisions should effectively cancel out, as long as they have 0 mean, and *ii*) since these are voluntary reviewers, we expect a majority of them to have at least some passion and knowledge about movies. Furthermore, this falls in line with the idea that while experts are more likely to spot quality, individually they are not necessarily unbiased. Each expert has his own preferences, and this matters given that they are a small group. Clearly, the important broader question of whether the unconditional expression of individual opinions, even from unexperienced consumers, is desirable – is an important area of work in the future. Here, we are just scraping the surface, although we believe some intriguing facts arise.

6 Conclusions and Future Research

In this paper we document a source of bias in online consumer generated ratings systems: better products are purchased by more knowledgeable consumers, who are more stringent. We think of this higher stringency as reflecting higher standards, or a higher ability to detect failures, or both. As a result, high products' relative reviews are unfairly low compared to those of their inferior alternatives. In other words: the better the product, the more its typical buyer expects from it, the more downward biased its ratings will be. This is exactly the opposite of what normatively desirable: reviews should help separate the good from the bad, not conflate the two.

We show that this is more than a theoretical possibility. Using movie ratings data scraped from a well known and commonly used website, we test our claims and find strong support for them. In line with our intuition, experienced consumers do indeed choose different

movies compared to their less experienced peers. In particular, they choose products that are intuitively more likely to be of high quality, for instance Academy Awards winners. Moreover, they rate them much more stringently: for some movies, enthusiastic recommendations from non-experienced consumers coexist with harsh expert ratings. What was surprising for us is that this appears to be true not only on average, but for virtually every movie in our sample, whether comedy or horror, popular or niche, old or new.

We then point out to the fact that, however sizeable this bias is, correcting for it is fairly straightforward. We do so by computing alternative ratings in a world in which experts and non-experts watch and review movies in fixed proportions: our newly generated ratings seem to perform better than the website ones in a variety of dimensions, and robustly in the parameters we employ. We still did not attempt to measure the welfare gains our proposed policy yields. However, we believe our analysis so far suggests these are sizeable. To be sure, we also believe than even the biased existing ratings are, at least for this specific market, much more desirable than no ratings at all, so that overall welfare effects, despite not approximating first-best, are still positive.

Though our paper deals with online product ratings, we believe the mechanisms we highlight to be considerably more general. One prominent example is US colleges grading policies. Moore et al. [2010] show that ratings inflation varies considerably across colleges, and that employers do not properly correct for this bias. As a result, students from more grade inflated institutions are unfairly advantaged over their peers experiencing stricter grading policies. It is worth emphasising that the stakes for employers are likely much higher than those of consumers in our data, and that colleges' grading policies are both transparent and widely debated.[19] Nevertheless, no correction takes place, offering further validation for our assumption of naïve learning.

We see a variety of potential avenues for future research, including a more detailed study of reputations dynamics over time, the relationship between popularity and reputation, the impact of social learning on market structure, and whether individual ratings purely reflect consumer satisfaction - as assumed here, for simplicity - or are instead the result of more complex psychological forces, or incentives. This would complement our present work, and we intend to carry such analysis in the near future.

[19]For example, grading inflation in the Ivy League has received considerable media attention. Moreover, some universities, e.g. Princeton, have been emphasising their stricter grading standards compared to some of their competitors (e.g. Harvard, Yale).

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Fast and slow learning from reviews. 2017.

Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

Omar Besbes and Marco Scarsini. On information distortions in online ratings. *Operations Research*, 2018.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.

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.

Luis Cabral and Ali Hortacsu. The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics*, 58(1):54–78, 2010.

Luis Cabral and Lingfang Li. A dollar for your thoughts: Feedback-conditional rebates on ebay. *Management Science*, 61(9):2052–2063, 2015.

Andrew Caplin, John Leahy, and Filip Matějka. Social learning and selective attention. Working paper, National Bureau of Economic Research, 2015.

Yeon-Koo Che and Johannes Horner. Optimal design for social learning. *Quarterly Journal of Economics*, 2015.

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.

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

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

Apostolos Filippas, John Joseph Horton, and Joseph Golden. Reputation inflation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 483–484. ACM, 2018.

Tristan Gagnon-Bartsch and Matthew Rabin. Naive social learning, mislearning, and unlearning. 2016.

John Horton and Joseph Golden. Reputation inflation: Evidence from an online labor market. 2015.

Bar Ifrach, Costis Maglaras, and Marco Scarsini. Bayesian social learning with consumer reviews. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):28–28, 2014.

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*, 121(4):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.

Jonathan Lafky. Why do people rate? theory and evidence on online ratings. *Games and Economic Behavior*, 87:554–570, 2014.

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

Dara Lee Luca and Michael Luca. Survival of the fittest: The impact of the minimum wage on firm exit. 2017.

Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper*, (12-016), 2011.

Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. *Management Science*, 2016.

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.

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.

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

Rani Spiegler. Bayesian networks and boundedly rational expectations. 2014.

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

Raluca Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. 2017.

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.

7 Tables

Table 1: Variables in Data

Variable Names	Variable Description
Reviewer Demographics	
prop_US	Proportion of raters self reporting as located in the US.
prop_female	Proportion of raters self reporting as female.
prop_aged_<29	Proportion of raters self reporting as less than 29 years old.
prop_aged_30 − 44	Proportion of raters self reporting as between 30 and 44 years old (inclusive).
prop_aged_45+	Proportion of raters self reporting as at least 45 years old.
Movie Demographics	
movie_year	Year movie was released.
movie_genre	Genre of movie.
oscar_nom	Dummy variable which takes a value of 1 if movie was nominated for an oscar.
Ratings	
num_reviews	Number of ratings submitted.
overall_mean_score	Arithmetic mean of all ratings submitted by reviewers.
(non)Top_1000_mean	Arithmetic mean of (Non-)Top 1000 reviewers.
overall_num_xx_stars	Count of all ratings submitted with rating value = xx.
(non)Top_1000_num_xx_stars	Count of ratings by (Non-)Top 1000 reviewers with rating value = xx.
prop_(non)Top_1000	Proportion of raters who are classified as (Non-)Top 1000 raters (as defined by the website) who submit a rating.

Table 2: Summary Statistics: Reviewer Demographics

	Top 250 Ranked Movies					
	Overall	Horror	Drama	Comedy	Action	Box Office
Number of Movies						
N	242	219	71	103	177	978
Total Number of Reviews (000s)						
Mean	443.71	59.72	4.58	10.82	111.47	32.82
Standard Dev	372.18	91.68	4.72	15.75	202.17	47.94
Proportion Female						
Mean	0.16	0.15	0.12	0.12	0.10	0.24
Standard Dev	0.06	0.06	0.14	0.09	0.07	0.13
Proportion Aged < 29						
Mean	0.42	0.31	0.52	0.47	0.49	0.39
Standard Dev	0.09	0.13	0.17	0.19	0.21	0.13
Proportion Aged 30-44						
Mean	0.46	0.50	0.41	0.42	0.41	0.45
Standard Dev	0.05	0.08	0.12	0.10	0.13	0.08
Proportion Aged 45+						
Mean	0.13	0.20	0.09	0.12	0.11	0.16
Standard Dev	0.06	0.12	0.08	0.12	0.12	0.09
Proportion US Raters						
Mean	0.28	0.36	0.25	0.21	0.26	0.29
Standard Dev	0.07	0.15	0.22	0.15	0.13	0.15

Table 3: Choice Heterogeneity: Proportion of Top 1000 Regression

	Dependent variable: prop_top_1000	
	(1)	(2)
movie_year	-0.001^{***}	-0.002^{***}
	(0.000)	(0.000)
prop_US	0.038	-0.101
	(0.403)	(0.075)
num_reviews	0.000^{***}	0.000
	(0.000)	(0.000)
oscar_nom	0.79^{***}	0.10^{***}
	(0.163)	(0.028)
Genre Controls	No	Yes
Observations	1790	1790
Pseudo R^2	0.29	0.32

Note: *p<0.1; **p<0.05; ***p<0.01
Estimation is via Fractionl Logistic Regression. Displayed values are Marginal Effects and Standard Errors of Marginal Effects. Constant is suppressed.

Table 4: Choice Heterogeneity: Aged 45+ Reviewers Regression

	Dependent variable:	
	Number of Aged 45+ Reviewers	
	(1)	(2)
movie_year	0.465^{***}	-70.903^{***}
	(0.046)	(6.021)
prop_US	-927.700^{***}	2031.787^{***}
	(92.492)	(446.189)
num_reviews	0.065^{***}	0.065^{***}
	(0.002)	(0.002)
oscar_nom	5208.044^{***}	3562.522^{***}
	(582.327)	(529.776)
Genre Controls	No	Yes
Observations	1790	1790
Adj R^2	0.93	0.93

Note: *p<0.1; **p<0.05; ***p<0.01
Estimation is via OLS. Standard errors are Eicker-Huber-White Robust Standard Errors. Constant is suppressed.

Table 5: Rating Heterogeneity: Difference in Means between Top 1000 and Non-Top 1000 Reviewers Regression

	Dependent variable:	
	Top_1000_mean - nonTop_1000_mean	
	(1)	(2)
movie_year	-0.004^{***}	-0.001
	(0.046)	(0.001)
prop_US	0.666^{**}	0.442
	(0.296)	(0.294)
num_reviews	0.000^{***}	0.000^{***}
	(0.000)	(0.000)
oscar_nom	0.520^{***}	0.590^{***}
	(0.046)	(0.052)
Genre Controls	No	Yes
Observations	1790	1790
Adj R^2	0.09	0.14

Note: *p<0.1; **p<0.05; ***p<0.01
Estimation is via OLS. Standard errors are Eicker-Huber-White Robust Standard Errors. Constant is suppressed.

Table 6: Rating Heterogeneity: Top 1000 and Non-Top 1000 Subsample Regressions

	Dependent variables:	
	Top_1000_mean	nonTop_1000_mean
	(1)	(2)
movie_year	-0.020^{***}	-0.019^{***}
	(0.001)	(0.001)
prop_US	-0.327	-0.768^{***}
	(0.203)	(0.231)
num_reviews	0.000^{***}	0.000^{***}
	(0.000)	(0.000)
oscar_nom	0.636^{***}	0.046
	(0.046)	(0.042)
Genre Controls	Yes	Yes
Observations	1790	1790
Adj R^2	0.09	0.14

Note: *p<0.1; **p<0.05; ***p<0.01
Estimation is via OLS. Standard errors are Eicker-Huber-White Robust Standard Errors. Constant is suppressed. Column (1) uses only Top 1000 reviews, whereas, Column (2) uses only non-Top 1000 reviews.

Table 7: Counterfactual: Original Website Ranking v. Debiased Ranking for Top 250 Overall Movies

	Dependent variables:
	Original Rank - Debias Rank
movie_year	-0.501^{***}
	(0.148)
prop_US	0.001^{***}
	(0.000)
num_reviews	-0.001^{***}
	(0.000)
oscar_nom	15.78^{***}
	(6.069)
Genre Controls	Yes
Observations	250
Adj R^2	0.09

Note: *p<0.1; **p<0.05; ***p<0.01
Estimation is via OLS. Standard errors are Eicker-Huber-White Robust Standard Errors. Constant is suppressed.

Table 8: Counterfactual: Correlations Between Original Rankings and Debias Rankings - $\gamma_E = (0, 0.5, 1)$

	Page Rank	**Top 1000 Rank** ($\gamma_E = 1$)	**Non-Top 1000 Rank** ($\gamma_E = 0$)	**50/50 Rank** ($\gamma_E = 0.5$)
Page Rank	1			
Top 1000 Rank	0.62	1		
Non-Top 1000 Rank	0.78	0.23	1	
50/50 Rank	0.76	0.96	0.43	1

Values are Spearman Rank coefficients.

8 Figures

Figure 1: Top 1000 versus Non-Top 1000 Mean Review Scores for Movies - Aggregate

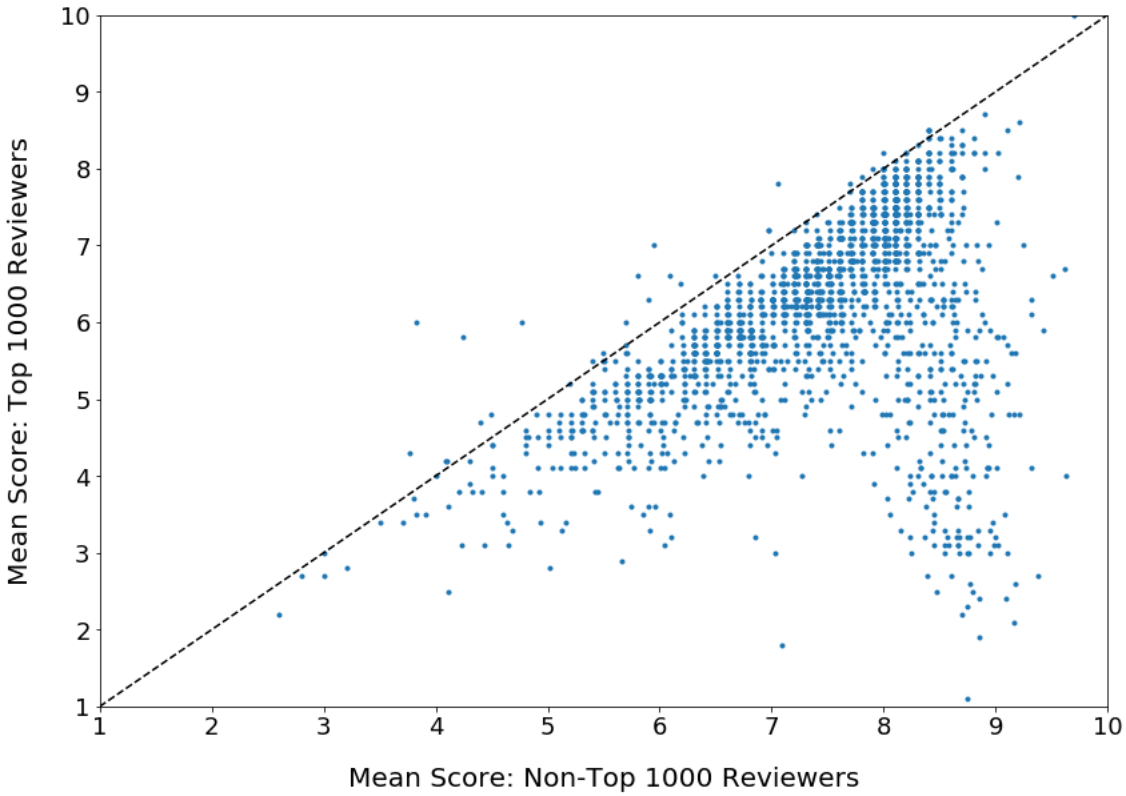

Figure 2: Top 1000 versus Non-Top 1000 Mean Review Scores for Movies - by Ranking

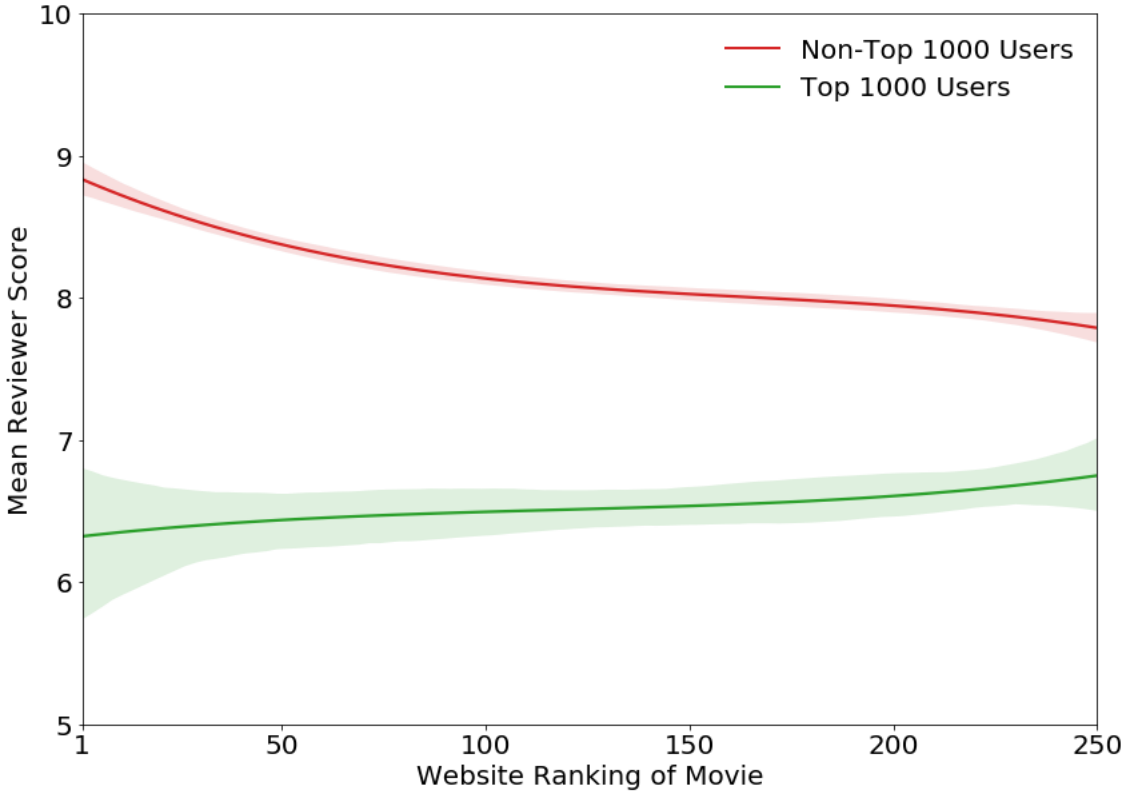

Figure 3: Top 1000 versus Non-Top 1000 Mean Review Scores for TV - by Ranking

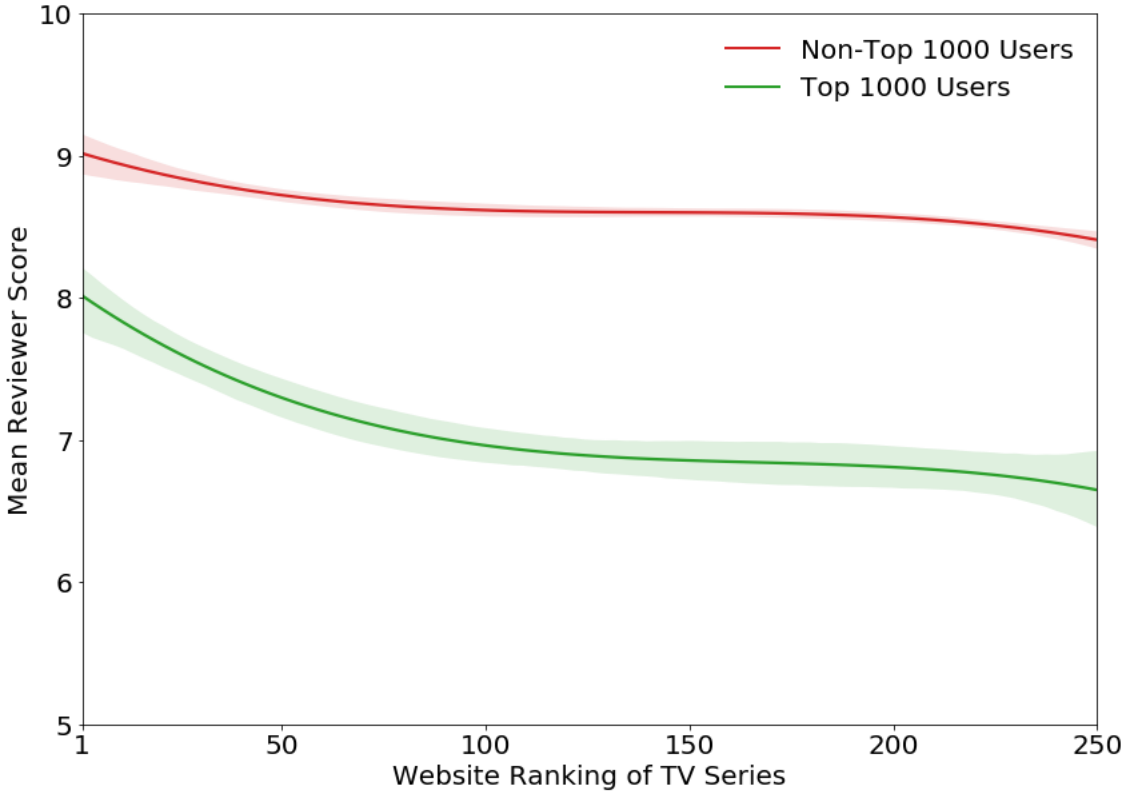

Figure 4: Top 1000 versus Non-Top 1000 Mean Review Scores for Movies - by Age

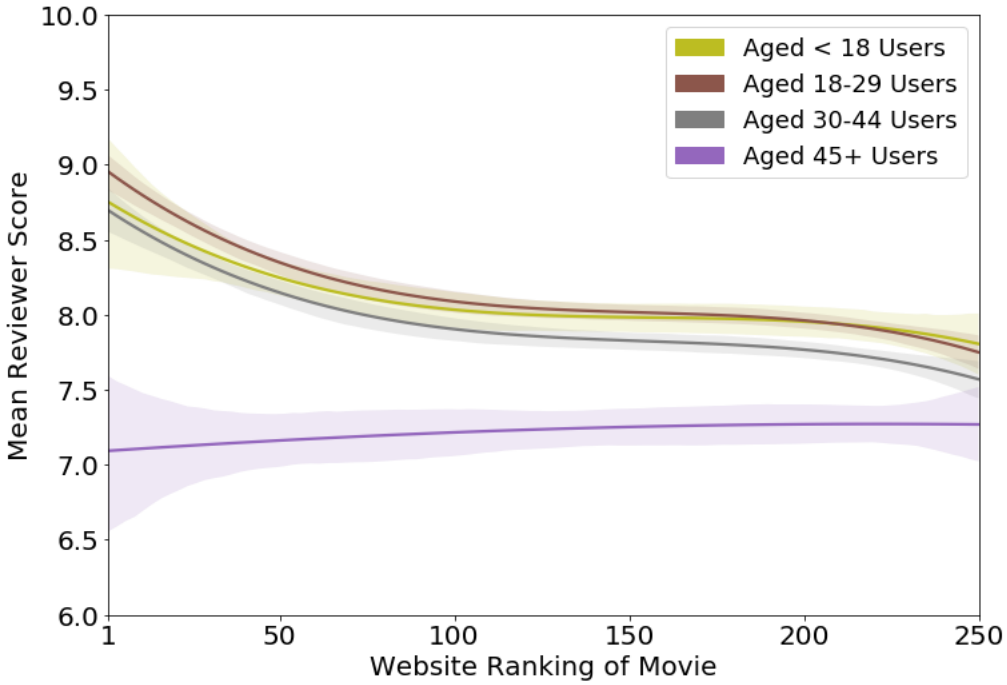

Figure 5: Top 1000 versus Non-Top 1000 Mean Review Scores for Movies - by Academy Award Nominations

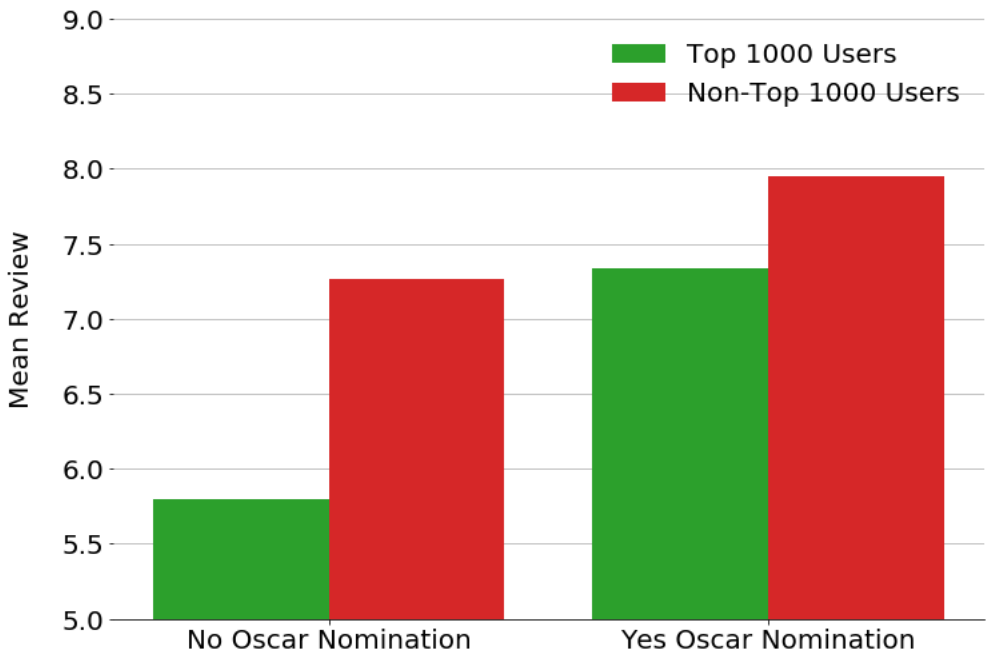

Figure 6: Distribution of Top 1000 and Non-Top 1000 Review Scores for Movies - Aggregate

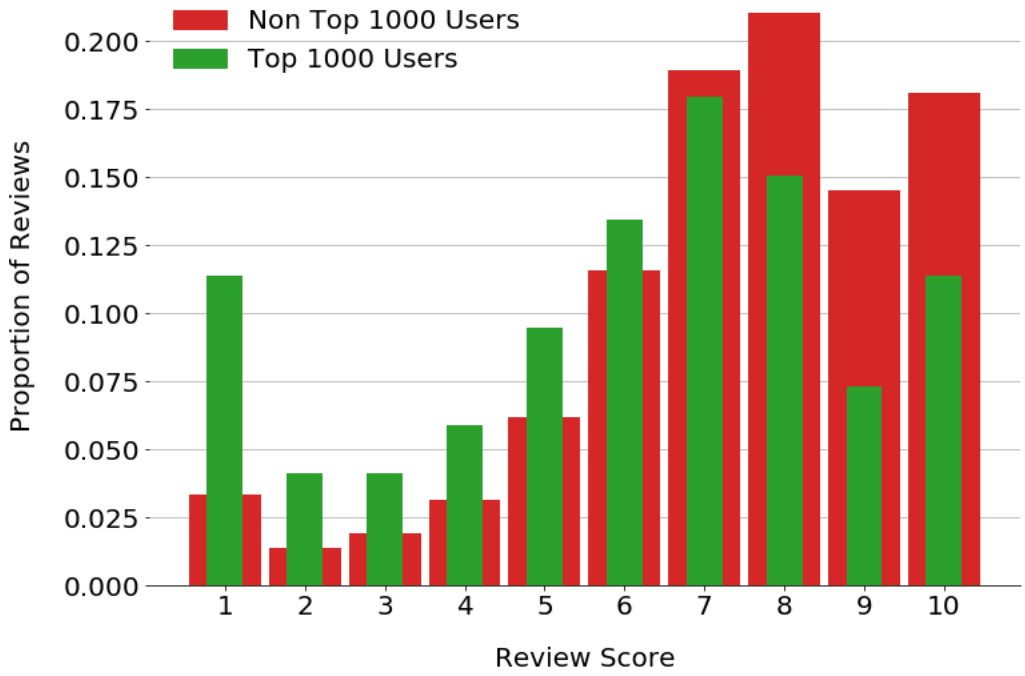

Figure 7: Bias Corrected versus Original Mean Review Scores - $\gamma_E = 0.5$

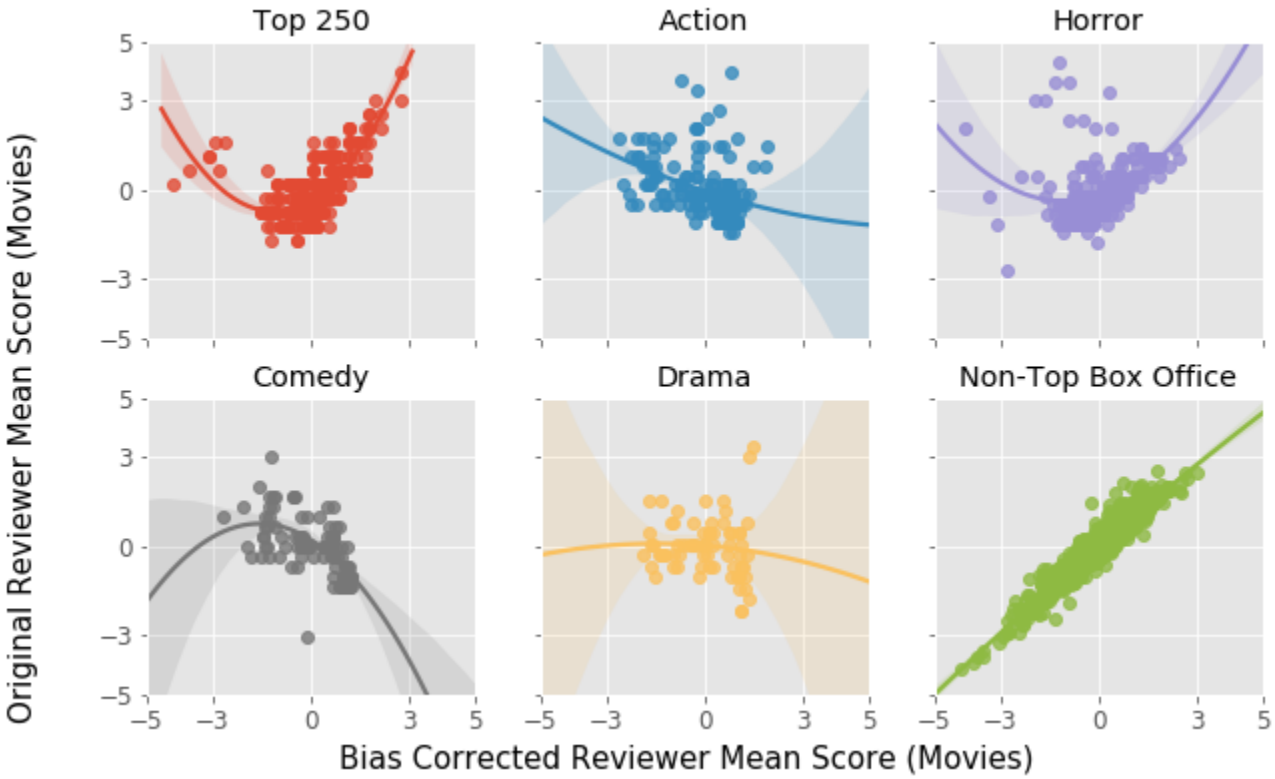

Figure 8: Contraction Ratio - $\gamma_E = 0.5$

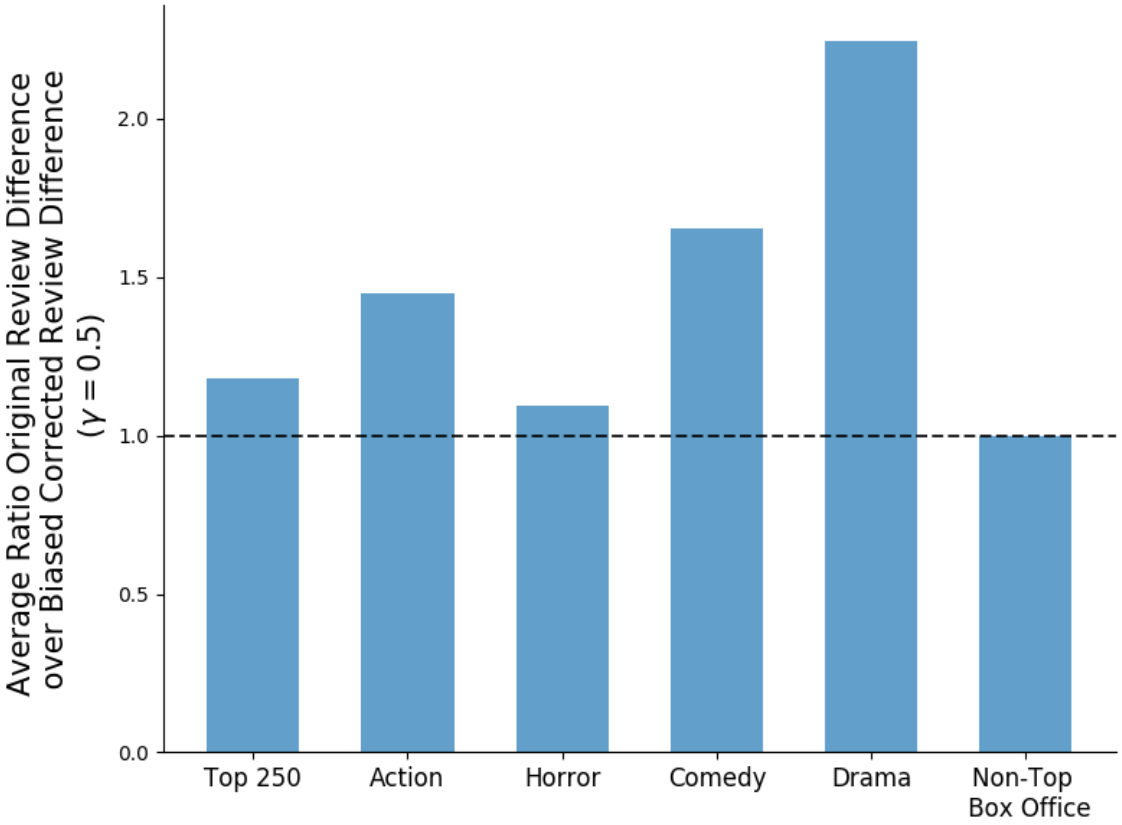

Figure 9: Rank Improvement Varying γ_E