Alone, Together: Product Discovery Through Consumer Ratings

Tommaso Bondi*
NYU, Stern School of Business

October 12, 2019

Please click HERE for the most recent version.

Abstract

Consumer ratings have become a prevalent driver of choice. I develop a model of social learning in which ratings can inform consumers about both product *quality* and their idiosyncratic *taste* for them. Depending on consumers' prior knowledge, I show that ratings relatively advantage lower quality and more polarizing products. The reason lies in the stronger positive consumer self-selection these products generate: to buy them despite their deficiencies, their buyers must have a strong taste for them. Relatedly, consumer ratings should not be used to infer which products are polarizing: what is polarizing *ex-ante* needs not be so among its buyers. I test these predictions using *Goodreads* book ratings data, and find strong evidence for them. *Goodreads* appears to serve mostly a matching purpose: tracking the behavior of its users over time reveals an increasing degree of specialization as they gather experience on the platform: they rate books with a lower average and number of ratings, while focusing on fewer genres. Thus, they become less similar to their average peer. Taken together, the findings suggest that consumer ratings contribute to both the long tail and, relatedly, consumption segregation. For managers, this illustrates, counterintuitively, the reputational benefits of polarizing products, particularly early in a firm's lifecycle, but only when paired with the ability to match with the right consumers.

*Email: tbondi@stern.nyu.edu. First and foremost, I thank Luís Cabral, John Horton and Raluca Ursu for their terrific support and guidance. I also thank Adam Alter, Alix Barasch, Xinyu Cao, Luca Citino, Daniel Csába, Fabrizio Dell'Acqua, Tulin Erdem, Apostolos Filippas, Sam Goldberg, Masakazu Ishihara, Gentry Johnson, Andreas Kraft, Xiao Liu, Alessandro Lizzeri, Yiangos Papanastasiou, Oren Reshef, Amin Sayedi, Vishal Singh, Daniel Stackman, Ryan Stevens, Joel Waldfogel, Jared Watson and Russell Winer, as well as seminar participants at NYU and Marketing Science (Rome) for insightful comments. Ruizhen Mai provided outstanding research assistance. I gratefully acknowledge financial support from the Centre for the Global Economy and Business at NYU Stern and the NET Institute. All errors are mine.

1 Introduction

Digitization has brought a substantial increase in variety in virtually all cultural markets: music, books, movies and TV shows are being produced at an unprecedented scale. In such a competitive landscape, in which thousands of products are fighting for consumers' attention (and money), it is of fundamental importance to understand how consumers sift through the large variety of products they are presented with.

The Internet also had a significant impact on how consumer discover products, in particular by means of consumer reviews. In the US, 95% of consumers have checked reviews in the last year, while 17% always check them before making a purchase (Brightlocal [2018]). Consumers use reviews to determine which products are objectively of high quality as well as to determine which products provide a good fit with their preferences. This inference, however, is made difficult by the fact that reviews include a subjective component precisely because, to some extent, such reviews measure specific consumer-product fit. In this context, how and what can consumers learn from peer-generated information?

This paper studies the nature and impact of consumer generated ratings in horizontally-differentiated products markets. Specifically, I develop a theoretical framework and identify a number of biases in peer-review systems due to consumer selection effects. I then apply the theoretical framework to the particular case of book reviews. I confirm the theoretical framework's empirical predictions, show that their size is considerable, and derive a number of additional stylized facts.

The first bias I identify is that differences in review scores understate differences in objective quality. This is due to the fact that high-quality products, by attracting a wider set of consumers (and thus reviewers), end up inducing reviews by buyers for whom the fit component of consumer satisfaction is relatively lower ("the curse of the best-seller"). In other words, the review system is biased against products with objective high quality: by attracting consumers with many different tastes, the product's success is also its curse.

Conversely, I show that consumer reviews favor "polarizing" products, that is, products whose fit component has very high variance (that is, consumers either love the product or hate the product). The idea is that, because of consumer self-selection, the fit component of reviews is very high: consumers for whom the fit component is low do not purchase the product and thus do not review it. In other words, for a given level of objective quality, niche products receive higher average ratings that general-interest products.

As a related point, I show that one cannot trust the variance of consumer ratings to accurately depict which products are polarizing (Clemons et al. [2006], Sun [2012]): because highly polarizing products are purchased by a homogenous set of buyers, their ratings will often have low dispersion. In fact, I show that this dispersion is often lower than it is for their best-seller counterparts, which attract a more diverse crowd.

To fix intuition, consider the following example. Suppose there are two competing alternatives of the same vertical quality: one polarizing (say, a far-right-wing political book), the other mainstream (say, a centrist book). The former will be bought by readers with right to far-right views. Assuming for simplicity that no other options are available, everyone else buys the centrist book. If we were to naïvely infer quality from ratings, the centrist book would appear relatively worse (because not all of its buyers adhere to its views) and relatively more polarizing (because its buyers are more diverse than its right wing alternative's).

I next extend the analysis to allow for learning. First, I show that biases about objective quality are self-correcting in nature: excessively high ratings today lead to a higher number of reviews from poorly-matched consumers, which in turn lowers the product's average review. Importantly, this correction is only partial, and long-run ratings display the same qualitative biases as short-run ones. Second, to the extent that consumers learn primarily about product fit, learning leads to divergence across consumers (and convergence within each consumer's basket). The opposite is true when consumers learn about vertical quality. In sum, whether dynamics imply collective convergence or divergence depends critically on the nature of learning, that is, whether consumers learn primarily about the objective quality of a product or rather about the product's fit to their taste.

In the second part of the paper, I test the above predictions using a newly-built dataset that combines information scraped from two book-related websites: *Goodreads* (the premier website for consumer-generated book ratings); and *Book Marks* (an aggregator of critics' reviews). In addition to aggregate statistics such as average rating and number of ratings, for each book I collect hundreds of time stamped individual ratings and full text reviews, together with information about the users (or critics) writing such reviews.

The empirical evidence is remarkably consistent with theory. Specifically, compared to *Book Marks*, *Goodreads* ratings contract quality differences and overvalue more polarizing products. Because, to some degree, critics self-select too, my empirical results understate the magnitude of each bias. Moreover, contrary to the premise of prior research, I find that the variance of ratings is not a good proxy of how polarizing a product is. Perhaps more revealing, the strongest predictor of high variance is a high number of ratings. In other words, as predicted by the theoretical framework, the first and second moments of product ratings are related: bestsellers, by attracting a more diverse set of buyers, receive a negative shock in terms of average reviews and a positive shock in terms of variance of reviews. Polarizing books, by contrast, receive a positive shock in terms of average reviews and a negative shock in terms of variance of reviews.

In order to examine the nature of social learning, I collect extensive *Goodreads* user data: for 1000 users, I observe a subset of their activity on the platform, including what they read, how they rate it, and what they intend to read in the future (as self-reported). I find that consumers become increasingly consistent with their choices, while less similar from their peers. In other words, consistent with the theoretical framework – in the particular case in which consumers learn about fit more so than quality – I find individual convergence and collective divergence.

Quantitatively, I find that, on average, consumers read 235% more obscure books in their last 50 *Goodreads* books compared to their first 50. Conversely, they read 69% fewer bestsellers.[1] The average and median number of ratings of books they experience are 81% and 93% lower when they get more experienced, respectively. Their genre choices are also more concentrated, while involving a 17% higher share of the most unusual genres. Moreover, in line with my model, I find that the average quality of books read by each individual (as proxied by average rating), (slightly) decreases over time.[2]

Because each consumer becomes more specialized, consumers look less alike as they get more experienced. The expected number of books and genres shared by two consumers go down by over by 83% and 16%, respectively. When each consumer is looking for good idiosyncratic matches, the level of agreement (and, thus, shared reads), naturally goes down. The same information, interpreted through the lenses of different tastes, leads to diverging consumption bundles.

Though a causal interpretation of these results is of course problematic — some consumers join *Goodreads* in order to find lesser known books that are a good fit for them to begin with — *Goodreads*'s sheer size compared to alternative platforms for book ratings and discovery (e.g., newspapers and literary blogs), on top of the nature of the information it provides (as studied in this paper), is likely to have had an important impact on both book discovery, evaluation, and consumption.

These findings are likely to generalize to other markets (cultural and not) in which product-consumer fit is plays an important role, and they inform publishers, platforms and

[1]Consumers read, on average 1.6 (18.4) and 5.5 (5.8) books in the bottom $10th$ (top $90th$) percentile of popularity when new and experienced on *Goodreads* respectively.

[2]Because more polarizing and lesser known books' average ratings are more upward biased, using them as proxy, if anything, understates the decline in quality.

consumers. Publishers should consider the more complicated tradeoffs of different design, pricing and advertising in contexts in which online word of mouth is key to long-term success, often opting for seemingly counterintuitive strategies. When thinking about product design, a naïve marketer might assume that a mainstream design maximizes the chances to satisfy buyers, thus minimizing the chances of negative word of mouth. However, a mainstream product might fail to attract a targeted, passionate crowd, and as a result obtain mediocre ratings, to the detriment of its long-term success.

When thinking about prices, the most natural assumption is that ratings might reflect the "quality of the deal" more so than quality *per se*.[3] This concern might be less relevant in markets – such as those for books or movies – in which prices are not as salient, or even fixed. In my model, high prices effectively work as a matching device: only consumers with a strong taste for the product will buy it. The opposite is true for low prices, which attract some consumers who do not have a strong taste for the product.[4]

On the platform side, I suggest an important channel through which *Goodreads* benefits Amazon – which acquired *Goodreads* in 2013 – at the expense of its competition. *Goodreads* (and similar platforms) influence the nature of product discovery and thus actively shape consumer choices, by allowing consumers to find lesser known products that match their taste, more so than simply identifying products of high quality.

Because the ability to provide consumers with little known titles is one of Amazon's key strengths, a shift towards lesser known products is key to its success. When selling bestsellers, competition is essentially on prices and shipping times (or immediate shelf availability). On the other hand, when selling little known books, the vast majority of Amazon's competition (rationally) rules itself out by not stocking them. Soon after Amazon's acquisition of *Goodreads*, the latter's growth spiked significantly. This contributed to the fragmentation of the book market, to Amazon's advantage: a more fragmented market advantages larger retailers.

For consumers, the key takeaway is to carefully consider the source of information they are presented with, and to be wary of simple aggregate statistics like the average and the variance of ratings. Product-consumers idiosyncrasies are as represented in ratings as the products' very characteristics. Particularly when using *Goodreads* to match with the right products, the key issue for consumers is to identify peers with similar taste: this way, idiosyncrasies in ratings would effectively reflect their future satisfaction.

[3] E.g., see Luca and Reshef [2019] in the context of *Yelp* restaurant ratings.

[4] This prediction is in line with empirical evidence on the perverse reputational effects of deep discounts, see for instance Byers et al. [2012].

The rest of the paper is structured as followed. Section 2 presents the related literature. Section 3 introduces the model. Section 4 describes the data. Section 5 my empirical strategy and findings. I conclude in Section 6. Appendices A, B, C, D and E contain the proofs, an additional model of quantity-quality tradeoffs in ratings, several model extensions, a detailed description of the data and the empirical results, respectively.

2 Related Literature

There is a currently very active and interdisciplinary literature focusing on the impact of digitisation on consumer choice and on broader market outcomes. Here, I am mostly going to discuss the subset of this research that focuses on online reviews. For exhaustive surveys, see Cabral [2012] and Tadelis [2016].

The majority of the existing literature on online reviews fall into two broad categories. The first line of research tries to quantify the causal impact of product reviews on sales, sales rank and prices. See Godes and Mayzlin [2004], Resnick et al. [2006] and Chevalier and Mayzlin [2006] for seminal contributions and Luca [2016] and Chen et al. [2017] for more recent ones. These papers consistently find that reviews are a key explanatory factor for products' success and lack thereof.[5]

The second stream studies the informational content of reviews. These papers have documented a variety of biases, due to social influence (Jacobsen [2015], Muchnik et al. [2013]), consumers' reciprocity towards sellers (Filippas et al. [2018]), sellers' manipulation (Luca and Zervas [2016]), or, most relevant for this paper, consumer self-selection (Li and Hitt [2008], Acemoglu et al. [2017], Vaccari et al. [2018], Besbes and Scarsini [2018]).

Vaccari et al. [2018] show, in line with this work, that with multiple products, ratings underestimate quality differences. However, the mechanisms are completely different. In their model, individual preferences are reference-dependent, so that high expectations are self-defeating. My results follow naturally from consumer trading offs between products' dimensions.

Acemoglu et al. [2017] consider a model of rational learning from reviews, and show that despite the complex and time-varying self-selection, a Bayesian learner can correctly infer the product's quality from ratings. Using a very similar sequential model, Besbes and Scarsini

[5] Also see Floyd et al. [2014], You et al. [2015], Watson et al. [2018] and citations there in for a discussion of how consumer trade off the average and the volume of ratings. Appendix B in my paper offers a complementary perspective.

[2018] show that accurate learning can be achieved under weak assumptions on consumer sophistication. However, they also point out that simply using the mean review as a proxy of quality leads to an incorrect long-run estimate.

Of these papers, only Vaccari et al. [2018] deals with multiple products. This is key, because ratings are increasingly employed to decide *what* – not *if* – to buy. Therefore, if all products' ratings are equally biased, *relative* ratings are accurate. Moreover, biases that are common to all platforms and products are easier for consumers to correct for: for instance, it is known that an average score of 4 out of 5 is good, but not stellar, on many platforms (Filippas et al. [2018]). Last, ratings are often used by platforms to form *rankings*: these determine not only consumer beliefs but also their considerations sets (Ursu [2018]).

When ratings' biases are systematic, consumer could, in principle, correct for them. This correction, however, is unlikely: De Langhe et al. [2015] document that consumers lack sophistication when interpreting reviews, and "navigate by the stars", even when they are likely to lead them astray. I will proceed with this behavioral assumption for most of this paper.

Another area of contribution of my paper is trying to infer product design from ratings. Do reviews tell us whether a product is polarizing? In influential work, Clemons et al. [2006] and Sun [2012] assume they do. Using a one product model, Sun [2012] defines polarizing products as those with higher "transportation costs"[6], and shows that – when consumers are uninformed about products' designs – their ratings have higher variance. Both my model and my empirical analysis question this assumption – in fact, I find that bestselling products are often associated with a higher variance of ratings.

My advice to platforms is partly at odds with some recent contributions to the literature on platform design and "crowdsourced exploration". Kremer et al. [2014], Papanastasiou et al. [2017], Che and Hörner [2017] and Vellodi [2018] consider the problem of a platform incentivising exploration by consumers, to generate positive externalities (through product discovery) and maximise their long-term utility. They show that, with rational consumers, the optimal policy involves a positive amount of spamming of new and unproven options. Ratings are modelled as unbiased signals of quality. While crowdsourced exploration is not the focus of this work, I show that the endogenous nature of ratings – which do not (just) reflect quality, and instead advantage lower quality and more polarizing options – might favour exploration even absent the platform's explicit intervention.

[6]Despite some *prima facie* differences, modelling niche products in terms of "transportation costs" is effectively equivalent to my theoretical approach.

It is interesting to contrast outcomes in environments in which consumers learn from the *opinions* of their predecessors with models of observational learning, in which they learn from their *actions*. A large literature, originating in Banerjee [1992] and Bikhchandani et al. [1992], has studied this setting.[7]

Models of observational learning are characterised by informational cascades, generating a "winner-takes-all" dynamic for sellers and an almost immediate breakdown in the aggregation of information (Zhang [2010], Tucker et al. [2013]).[8] In my model, the opposite happens: niche options are overvalued at the expenses of more popular ones, increasing market fragmentation. Not only information needs not lead to cascades, it might actually end them: when information about product fit is gradually revealed, consumers weight it more, causing their consumption bundles to become increasingly divergent.

Clemons et al. [2006], Dellarocas and Narayan [2007] and more recently Hosanagar et al. [2013] all investigate the distributional effects of platform recommendations, whether in the form of consumer ratings or personalized recommendation systems.[9] Hosanagar et al. [2013] show that recommendation systems yield more similar consumption bundles across consumers.

On the contrary, I show that consumers' consumption bundles diverge over time, as they discover increasingly lesser known products. Relatedly, Clemons et al. [2006], focusing on the beer industry, show that reviews are key to increase niche products' market shares.[10]

Important precedents studying the impact of reviews specifically in the book market include Chevalier and Mayzlin [2006], Sun [2012], Dobrescu et al. [2013] and Kovács and Sharkey [2014].[11] In line with one of my theoretical predictions, Kovács and Sharkey [2014] consider the consequences of a popularity shock for books. They demonstrate that upon

[7] Importantly for this study, Smith and Sørensen [2000] is the first to model observational learning in contexts with taste heterogeneity.

[8] There are limits to this, particularly in strategic environments with dynamic pricing, as recently studied by Sayedi [2018].

[9] Berman and Katona [2018] consider the related problem of personalisation versus exploration in the context of social media, and show that more refined curation algorithms need not create an echo chamber effect. The reason, they argue, lies in the endogenous users response: as the platforms becomes better at filtering out offensive and irrelevant information, users are incentivised to experiment more with whom they follow and link to.

[10] Clemons et al. [2006] also interpret a high variance in ratings as evidence of the product being polarizing, and show that it is correlated with higher sales. My model and data demonstrate that, in fact, a reverse causality problem might be present: higher sales lead to a more heterogeneous buyers base, which is in turns reflected by a higher variance in ratings.

[11] There is a broader literature on the impact of reviews in cultural markets. See for instance Dellarocas and Narayan [2007], Chintagunta et al. [2010], Moretti [2011], Cabral and Natividad [2016] and Gilchrist and Sands [2016] in the context of movies.

receiving a major literary award, winners receive many more, but on average worse, ratings than runner ups. Consistent with my model, they show that this effect can be traced to an increase in heterogeneity among the winners' buyers.

3 Theoretical Framework

3.1 A Baseline Model of Ratings

The baseline model in this Section features two competing products, $I = \{1, 2\}$, and a continuum set of buyers, $\mathcal{J}$. The products are both vertically and horizontally differentiated. While all consumers agree on the vertical dimension, the same is not true for the horizontal one, which represents idiosyncratic taste. Products differ in quality, price and design. Design measures how polarizing (or niche) the product is: a mainstream (high mass appeal, H) design will be inoffensive to all consumers, while a niche, or polarizing (low mass appeal, L) design will polarize consumers, who will either love it or hate it.

Throughout the paper, I will abstract from sellers' strategic decisions, and will instead focus on consumer beliefs, choices and ratings.[12]

Consumer j's utility for product i is given by

$$U_{ij} = Q_i + \theta_{i,j} - P_i.$$

$\theta_{i,j}$ represents idiosyncratic consumer-product match, and is drawn from a continuous and smooth cumulative distribution $F_{s_i}(\cdot)$ with mean 0, *iid* across consumers. Producers can select a product design $s_i \in \{s_L, s_H\}$, that is, they can modify the shape of $F_{s_i}(\cdot)$, subject to the constraint that its mean be fixed at 0. More specifically, following Johnson and Myatt [2006], designs are ranked in terms of demand rotations.

Definition 1. *We say that $F_{s_i'}(\cdot)$ is a rotation of $F_{s_i}(\cdot)$ if there exists a $\theta^\dagger_{s_i}$ such that*

$$F_{s_i}(\theta) < (>) F_{s_i'}(\theta) \iff \theta < (>) \theta^\dagger_{s_i}.$$

Intuitively, $F_{s_i'}(\cdot)$ concentrates more mass around $\theta^\dagger_{s_i}$ than $F_{s_i}(\cdot)$ does. In economic terms, this means that the seller can decide between more mainstream designs, which are moderately

[12]Notice that in markets such as the one for books, moral hazard concerns are arguably limited. Publishers and writers compete mostly on product positioning, and perhaps pricing.

appealing to most consumers and offensive to none, and more niche ones, which will be loved by some, loathed by others.

The following result, again due to Johnson and Myatt [2006], allows me to restrict attention on these two extreme designs, which simplifies the notation, though it should be noted that my results hold more generally.

Lemma 1. *When the family of distributions is ordered by a sequence of rotations, and the corresponding sequence of the $\theta^{\dagger}_{s_i}$'s is decreasing, then profits are quasi-convex in design, leading firms to pick one of two extremes: niche (L) and mainstream (H).*

Contrary to Johnson and Myatt [2006], I assume $\mathbb{E}_s(\theta) = 0$, for every design s. This is in line with the interpretation of θ as consumer-idiosyncratic taste. It also allows me to think of demand rotations in terms of second order stochastic dominance: more niche designs are dominated (in the second order stochastic sense) by less niche ones. Last, I make one natural assumption.

Assumption 1. *For fixed prices and qualities, the mainstream product's market share is greater than or equal to the niche product's:*

$$M_H := P(\theta_H > \theta_L) \geq M_L := 1 - M_H.$$

Throughout this Section, I assume that prices are always observable to all consumers, while both vertical quality and their match value for each product are potentially unknown. In Section 3.2, they are both the object of social learning.

Upon choosing a product, each buyer reviews it subjectively, by reporting her own experienced utility.[13] That is, we have

$$\mathcal{R}_{ij} = \begin{cases} Q_i + \theta_{ij} & \text{if } i \in argmax\{\mathbb{E}(U_{1,j}), \mathbb{E}(U_{2,j})\}, \\ \varnothing & \text{otherwise.} \end{cases}$$

Truth-telling is a sensible assumption on platforms – like *Goodreads* – that motivate consumers to leave product reviews at least partly to receive future personalized recommendations. Denote by $\mathcal{J}_1$ and $\mathcal{J}_2$ the sets of buyers of product 1 and 2 respectively. That is,

[13]Note this buries two assumptions: that the consumer always reviews the chosen product, and that such review is truthful, that is, it purely reflects the consumer's utility. I discuss the robustness of my results to alternative, data motivated modelling choices, such as extremity bias and motivated ratings, in Appendix C.

$$\mathcal{J}_1 = \{j \in \mathcal{J} \mid \mathbb{E}(Q_1) + \mathbb{E}(\theta_{1,j}) - P_1 \geq \mathbb{E}(Q_2) + \mathbb{E}(\theta_{2,j}) - P_2\},$$

and similarly for $\mathcal{J}_2$. The expectations depend on consumers' knowledge of the product attributes, as discussed below. Moreover, denote by $F_{s_i}^{\mathcal{J}_i}$ the conditional distributions of θ_i, $i = 1, 2$:

$$F_{s_i}^{\mathcal{J}_i}(\theta_i) = \int f_{s_i}(\theta_{i,j}) d\mathcal{J}_i, \quad j = 1, 2.$$

Denoting by $G_{\mathcal{R}_i}(\cdot)$ the CDF of product i ratings, it is immediate to see that $G_{\mathcal{R}_i}(\cdot)$ satisfies

$$G_{\mathcal{R}_i}(Q_i + \theta_i) = F_{s_i}^{\mathcal{J}_i}(\theta_i).$$

I am interested in studying the properties of the mean and variance of $\mathcal{R}_i$, which are given by

$$\mathbb{E}(\mathcal{R}_i) = \int \mathcal{R}_{ij} d\mathcal{J}_i = Q_i + \mathbb{E}_{F_{s_i}^{\mathcal{J}_i}}(\theta_i), \qquad Var(\mathcal{R}_i) = Var_{F_{s_i}^{\mathcal{J}_i}}(\theta_i).$$

Note that while *prima facie* Q_i only shifts up the distribution of ratings, and thus does not enter $Var_{F_{s_i}^{\mathcal{J}_i}}(\theta_i)$ directly, it does so indirectly through consumer self-selection, since $\mathcal{J}_i = \mathcal{J}_i(Q_i, Q_{-i}, s_i, s_{-i}, P_i, P_{-i})$.

Before describing a sequential model of social learning in Section 3.2, I characterise the features of product ratings in four scenarios: *i*) consumers ignore both quality and fit, *ii*) consumers only know quality, *iii*) consumers only know their fit with each product and *iv*) consumers know both quality and fit. As I will show, the nature of *ex-ante* information possessed by consumers shapes both the direction and the magnitude of the biases we observe in ratings.

3.1.1 Two Dimensional Uncertainty

When consumers ignore both quality and fit attributes, assuming $P_1 = P_2$ they will choose randomly, and thus ratings accurately reflect their *ex-ante* distribution. Therefore, $\mathbb{E}(\mathcal{R}_i) = Q_i$ and $Var(\mathcal{R}_i) = Var_{F_{s_i}}(\theta_i)$, $i = 1, 2$.

In other words: because ratings biases in my model follow from patterns of consumer self-selection, when consumer know nothing – and thus, choose randomly – ratings are unbiased. While mathematically trivial, this fact has important implications for platform design: sometimes, perhaps counterintuitively, platforms might have incentives to *censor* consumer ratings as consumers become more informed about products.

3.1.2 Quality Uncertainty

In this case, consumers choose solely based on products' prices and their specific taste matches. For example, consumers might know the genre of each book, but ignore their quality. The next Proposition is one of the paper's central results.

Proposition 1 (Ratings with Quality Uncertainty)**.** *When consumers only choose products based on their designs,*

- *If* $s_1 = H$, $s_2 = L$,
$$\mathbb{E}(\mathcal{R}_1) - \mathbb{E}(\mathcal{R}_2) < Q_1 - Q_2.$$
Moreover, $Q_1 > Q_2$ *does not imply* $\mathbb{E}(\mathcal{R}_1) > \mathbb{E}(\mathcal{R}_2)$, *and* $Var(\theta_2) > Var(\theta_1)$ *does not imply* $Var(\mathcal{R}_2) > Var(\mathcal{R}_1)$.

- *When* $s_1 = s_2$, *ratings are upward biased*[14]*:* $\mathbb{E}(\mathcal{R}_i) > Q_i$, $i = 1, 2$, *but relatively fair:*
$$\mathbb{E}(\mathcal{R}_1) - \mathbb{E}(\mathcal{R}_2) = Q_1 - Q_2.$$

Moreover, the products look equally polarizing *ex-post*: $Var(\mathcal{R}_2) = Var(\mathcal{R}_1)$.

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

To gather some intuition for the first result, assume we knew that the mainstream product was chosen. Since its valuation among consumers is fairly concentrated, this was likely caused by a distaste for the niche alternative. So, the mainstream product's ratings will not be particularly upward biased, and will reflect the opinions of a diverse set of consumers (namely, everyone but those with a high θ_2).

On the flip side, when observing a consumer choosing the niche product, it will be relatively more likely this is due to a strong taste for it than to a (statistically rare) distaste for the mainstream alternative. Reviews of niche products reflect the opinions of their fans, while those of mainstream products reflect the opinions of anyone who is not a fan of the available alternatives. This implies stronger matches – and thus more upward-biased ratings – for the niche product.

[14]With only two products, it is theoretically ambiguous whether this upward bias is larger when both products are mainstream or niche; with a large number of products, however, the ambiguity disappears and consumers prefer a niche-heavy market. This is because as the number of products grows, they are almost guaranteed to find a near-perfect match. Such high values for θ are much more unlikely for mainstream products. Moreover, it can be shown that in markets with many options, all sellers apart from the highest quality ones will opt for a niche design (Johnson and Myatt [2006], Bar-Isaac et al. [2012]).

Note that my result does not depend on the mainstream product capturing a larger market share (for instance due to skewness in the distribution of F_{s_L}). That is, Proposition 1 is *not* saying "The niche product sells only to a small set of fans, so it obtains fewer but better reviews". The result holds even when the two market shares are equal. A more correct interpretation would be "Every recommendation system in which raters self-select based on product fit to maximise their utility relatively advantages more polarizing options".

It is worth emphasising that the result also does not depend on competition, so that it is applicable to the much more studied setting of online reviews in monopoly (e.g., Acemoglu et al. [2017], Besbes and Scarsini [2018], Papanastasiou et al. [2017]).

Corollary 1. *Assume each consumer is choosing between a product and an outside option of quality c. When c is high enough, the product's ratings are increasing in its nicheness, for given quality.*

Example For concreteness, assume $\theta_1 \sim N(0,1)$ and $\theta_2 \sim N(0,4)$. Then, if consumers were choosing purely based on their taste match with each product, market shares would be symmetric:

$$M_1 = P(\theta_1 \geq \theta_2) = P(-\theta_1 \leq -\theta_2) = P(\theta_1 \leq \theta_2) = M_2,$$

where the second to last equality is due to symmetry. Thus, $M_1 = M_2 = \frac{1}{2}$. Despite this, the *conditional* match quality is asymmetric. The average match between the mainstream (niche) product and its buyers is given by $\mathbb{E}(\theta_1|\theta_1 > \theta_2)$ ($\mathbb{E}(\theta_2|\theta_2 > \theta_1)$). We have

$$\mathbb{E}(\theta_1|\theta_1 > \theta_2) \approx 0.35 < 1.42 \approx \mathbb{E}(\theta_2|\theta_2 > \theta_1).$$ [15]

Note that $P(\theta_1 > 0.35) = 0.36$, $P(\theta_2 > 1.42) = 0.23$. That is, if only one representative consumer for each product were to leave ratings, the consumer of the mainstream product would like it more than 64% of his peers, while that figure goes up to 77% for the niche one. The latter is roughly twice as removed from the (product-specific) median (and mean, by symmetry) consumer as the former. In other words: while both ratings are upward biased, the bias is larger for niche product.[16]

[15] When comparing $\mathbb{E}(\theta_1|\theta_1 > c)$ and $\mathbb{E}(\theta_2|\theta_2 > c)$, one can directly appeal to the well known formula for truncated normals:

$$\mathbb{E}(N(\mu, \sigma^2) \mid N(\mu, \sigma^2) > c) = \mu + \sigma \frac{\phi(c)}{\Phi(c)},$$

which is straightforwardly increasing in σ for every μ, σ and c.

[16] More generally, given $\theta_1 \sim N(0, \sigma_1^2)$ and $\theta_2 \sim N(0, \sigma_2^2)$ with $\sigma_1^2 < \sigma_2^2$, the following are true: *i)* $\mathbb{E}(\theta_1|\theta_1 > \theta_2) < \mathbb{E}(\theta_2|\theta_2 > \theta_1)$, *ii)* $P(\theta_1 > \mathbb{E}(\theta_1|\theta_1 > \theta_2)) > P(\theta_2 > \mathbb{E}(\theta_2|\theta_2 > \theta_1))$, *iii)* $\frac{\partial \mathbb{E}(\theta_i|\theta_i > \theta_{-i})}{\partial \sigma_{-i}} < 0$,

Last,

$$Var(\theta_1|\theta_1 > \theta_2) \approx 0.87, \quad Var(\theta_2|\theta_2 > \theta_1) \approx 1.96.$$

Both products look less polarizing than they are – a consequence of the fact that their most negative ratings are missing – but the effect is stronger for the more polarizing product 2:

$$\frac{Var(\theta_2|\theta_2 > \theta_1) - Var(\theta_2)}{Var(\theta_2)} = \frac{4 - 1.96}{4} = 0.51 > 0.13 = \frac{1 - 0.87}{1} = \frac{Var(\theta_1|\theta_1 > \theta_2) - Var(\theta_1)}{Var(\theta_1)}.$$

That is, while consumer self-selection cuts the *ex-ante* variance essentially in half for product 2, the decrease is only 13% for product 1. I will expand on this idea in Section 3.1.4, and show that under certain assumptions, a stronger version of this result holds: the mainstream products' ratings can in fact display a *higher* variance, reflecting its more heterogeneous buyers pool.[17]

3.1.3 Fit Uncertainty

In some settings, consumers have a good idea of quality and are mostly concerned about their fit. For example, they might know that one writer is talented, but they are not sure if they are going to like the genre of her newly released book; or be aware a restaurant obtained a Michelin star, but suspect it is too spicy for them, and so forth.

The first thing to notice is that this form of learning is relatively more important for niche products: because their designs are more polarizing, avoiding mismatches is relatively more of a concern for consumers than maximising quality. I find striking empirical evidence for this hypothesis empirically in Section 4.[18]

First assume consumers are in the dark regarding their idiosyncratic match values with each product: for example, they are exploring a new market and are unable to figure out how their taste map into that of different reviewers. In this case, their choices will be based purely on quality and price: they will choose product 1 if and only if $Q_1 - P_1 \geq Q_2 - P_2$. Because consumers are not self-selecting according to their taste, we have the following:

$\frac{\partial \mathbb{E}(\theta_i|\theta_i>\theta_{-i})}{\partial \sigma_i} > 0$ for $i = 1, 2$ and $iv)$ $\frac{\partial P(\theta_i>\mathbb{E}(\theta_i|\theta_i>\theta_{-i}))}{\partial \sigma_{-i}} > 0$, $\frac{\partial P(\theta_i>\mathbb{E}(\theta_i|\theta_i>\theta_{-i}))}{\partial \sigma_i} < 0$ for $i = 1, 2$.

[17] The Assumptions are sufficient but not necessary. The initial example I offered does not satisfy them, and still displays a reversal between *ex-ante* and *ex-post* variances.

[18] It is important to clarify a conceptual difference with previous work. The type of horizontal learning I describe here has little to do with consumers making inference of which product is more polarizing. In fact, when consumers are risk neutral, that information alone provides no value, since both shocks have mean 0. Learning here takes the form of making inference on the idiosyncratic match value θ each product will provide.

Proposition 2 (Ratings with Fit Uncertainty)**.** *When consumers know products' quality, but not their fit with each product, both the average and the variance of ratings are unbiased:*

$$\mathbb{E}(\mathcal{R}_i) = Q_i, \qquad Var(\mathcal{R}_i) = Var(\theta_i).$$

3.1.4 Ratings with Two Dimensional Information

Last, I consider a scenario in which consumers possess some information over both products' attributes. For example, consumers might be aware that a writer just won an important award *and* favor his genre.

For the sake of simplicity, to isolate the role of quality differences I start by considering the situation in which the two products have the same design and price. We have the following result - together with Proposition 1, a central one.

Proposition 3 (Ratings with Informed Consumers)**.** *Assume $Q_1 > Q_2$ and that $s_1 = s_2$. Then, ratings accurately rank products: $\mathbb{E}(\mathcal{R}_1) > \mathbb{E}(\mathcal{R}_2)$. Nevertheless, they relatively advantage the lower quality product, that is, they underestimate quality differences:*

$$\mathbb{E}(\mathcal{R}_1) - \mathbb{E}(\mathcal{R}_2) < Q_1 - Q_2.$$

The Proposition shows that ratings understate the relative quality of the better product. The Proof builds on the following fact:

$$\mathbb{E}(\theta_1 \mid Q_1 + \theta_1 > Q_2 + \theta_2) < \mathbb{E}(\theta_2 \mid Q_2 + \theta_2 > Q_1 + \theta_1).$$

That is, since product 2 is vertically worse, the mean idiosyncratic taste shock for its buyers is higher. One way to see this is that higher vertical quality / lower price can persuade buyers to go with a product even though *it is not the best match for them*. That is, there will exist a non-empty set $\underline{\mathcal{J}_1} \subset \mathcal{J}_1$ such that for every $j \in \underline{\mathcal{J}_1}$ we have $Q_1 + \theta_{1,j} > Q_2 + \theta_{2,j}$ even though $\theta_{1,j} < \theta_{2,j}$. Though individually rational, these decisions imply that, on average, consumers are better matched with the vertically inferior product. This favors the relative reputation – but hurts the market share – of the inferior product compared to the case in which the two products were vertically indistinguishable.

This is the case of the very high quality noir book which attracts all sorts of readers, part of whom will, nonetheless, be unimpressed with the plot twists and find the book too dark. It is important to stress that consumer choices are rational: each consumer is buying the subjectively optimal option. However, the higher quality option will face the higher burden of proof of satisfying consumers for whom it is not a perfect match. Here, *more ratings means harsher ratings.*

Notably, I have analyzed the effects of products' asymmetries on reputation one by one. In reality, products simultaneously differ in design, quality and price, and their (relative) ratings will reflect each of these dimensions. Because quality and price enter the utility function equally, and are both vertical, their role on ratings are opposite. While a higher quality yields higher market shares and (thus) lower ratings, a higher price achieves exactly the opposite: lower market shares but high ratings, coming from consumers whose taste of the product was high enough to tolerate its high price. Price therefore operates as a matching device.

How do product characteristics interact in a competitive market? That is, are high or low quality products more likely to have a niche design? Johnson and Myatt [2006] and Bar-Isaac et al. [2012] both show analytically, in related settings, that the unique equilibrium features a cutoff strategy for firms, who choose mainstream designs only when their quality is above a certain threshold. This is intuitive: a niche design is effectively a differentiation tool when competing on the vertical dimension is prohibitive. However, for the sake of ratings, in light of Propositions 1 and 3, it compounds the two biases: high quality mainstream products are bought by everyone without a (really) strong taste for one of the lower quality, niche alternatives.

Next I turn to the question of how correctly consumer ratings describe products' designs when consumers self-select on both the quality and the taste dimension. A straightforward conjecture is that, since product designs drive the variance of the taste shocks, a more polarizing design should correspond to a higher variance in ratings.

In influential work, Clemons et al. [2006] and Sun [2012] take exactly this approach. They then demonstrate the effect of the variance of ratings of Amazon books: a high variance increases sales only when the average rating is low. They interpret this as low quality products still being able to attract a well matched crowd when polarizing enough.

However, I show (theoretically and empirically) that a product's ratings' variance needs not be indicative of its design, at least in the case in which consumers can infer the product's fit before purchasing it (e.g. stance of political books, type of a restaurant's cuisine, genre of a movie, ...). This can translate in a complete reversal of *ex-ante* and *ex-post* variances, as shown in the following Proposition.

Proposition 4. *When quality differences are large enough, the niche product ratings have lower variance than the mainstream product.*

The reason for this reversal is that all of those who did not like the niche product's design are not buying (and thus rating) it. For the mainstream product, however, this selection is much weaker, and thus its buyers come from a more diverse group, often resulting in more diverse ratings[19].

Analytically, notice that this finding results from the combination of two forces: on one hand, the niche product has higher unconditional variance. On the other hand, we can make more inference on θ_L conditional on choice than we do on θ_H, so that the *reduction* in variance is larger for the niche product.

One reason why a high variance might help products with low average ratings is that *good news are more persuasive when they are precise; bad news are less damaging when they are noisy* (Harbaugh et al. [2016]). They demonstrate this in the context of firms' financial reports. It is possible that highly spread ratings help products with a low average rating not by informing some consumers of their match value, but rather by decreasing the ratings' credibility, and thus harmfulness. Importantly, this argument does not hinge on product design and matching.

3.1.5 Combining the Four Cases

Summing up the findings from the previous four informational environments:

- **No Information**: Consumers choose randomly.
- **Only Qualities Known**: The market is concentrated around the vertically better option. Each product's ratings reflect both its quality and its design.
- **Only Match Values Known**: Both products' ratings are upward biased. If the products have the same design, ratings accurately reflect quality and variance differences. If product 1 is more mainstream than product 2, then *i*) its average ratings are relatively lower biased (and possibly lower in absolute terms), and *ii*) product 2's ratings might display a lower variance.
- **Both Qualities and Match Values Known**: Average ratings always relatively advantage the lower quality product. This is a consequence of consumers trading off the vertical and the horizontal dimensions. The same is true for the variances.

[19] Goodreads' own ranking for the most polarizing books of all times clearly shows the consequences of wrongly interpreting the information contained in the variance of ratings: `https://www.goodreads.com/list/show/6199.The_Most_Polarizing_Books_Of_All_Time`.

The biases are exacerbated (alleviated) when the higher quality product has a more mainstream (niche) design.

3.2 Social Learning

The previous analysis aimed to highlight systematic biases in ratings as a function of consumers' private information. This information was taken as exogenous. While the simplified framework allowed me to more starkly illustrate the presence and magnitude of ratings biases, it also left out one key components: social learning from consumer reviews. In this Section, I incorporate it, and show that consumers' beliefs and choices crucially depend on both the presence and the nature of social learning.

Notation The following will ease exposition going forward: define ξ_i the distribution followed by $\theta_i - \theta_{-i}$, or in other words, the relative taste preference for product i. I therefore denote by $F_{\xi_i}(\cdot)$ the CDF of $\theta_i - \theta_{-i}$, and by $f_{\xi_i}(\cdot)$ its density. Moreover, since only *relative* qualities matter, denote by Q_i the quality *advantage* of product i and, without loss of generality, let $Q_1 > 0 > Q_2$.

The timing for this model is as follows.

Period 0 Two products, with exogenously fixed prices $P_i \geq 0, i = 1, 2$ and designs $s_i \in \{L, H\}, i = 1, 2$ are offered on the market.

Period 1 A continuum of consumers, each privately (un)informed about the products' characteristics, enters the market, and purchase their preferred option given their private information and the posted prices.

Period 1.5 Every Period 1 consumer reviews the product she purchased truthfully.

Period 2 A mass one of consumers enters the market, each purchasing their preferred option given the products' ratings and prices.

$\vdots$

Period t A mass one of consumers enters the market, each purchasing their preferred option given $t-1$ generation's products ratings and prices.

Period t.5 Every Period t consumer reviews the product she purchased truthfully.

$\vdots$

What does "given product ratings" mean? First, I have to specify what the object of learning is. Then, I have to formalisz the learning process. In line with recent work – see for instance Acemoglu et al. [2017] – I will contrast social learning by *rational vs naïve consumers*.

The key difference between these two groups is that rational consumers perfectly understand the self-selection that occurred in the previous generations of consumers, and thus ratings, and correct for it. Naïve consumers, on the other hand, take ratings at face value, without accounting for the predictable biases contained therein. Since in my model reviews contain bias, but no noise[20], rational social learners (RSL) are effectively able to infer products characteristics perfectly upon observing reviews.

3.2.1 Learning about Quality

In this scenario, consumers make inference about quality given the ratings left by the previous generation. Given the noiseless nature of ratings, rational consumers can perfectly infer quality: denoting by $\mathbb{E}^t_{RSL}(\cdot)$ their expectations, we have

$$\mathbb{E}^t_{RSL}(Q_i|\mathcal{R}_i^{t-1}, \mathcal{R}_{-i}^{t-1}) = Q_i, \quad \forall\ Q_i, s_i, P_i, Q_{-i}, s_{-i}, P_{-i}.$$

Naïve social learners (NSL), on the other hand, fail to recognise the selection bias incorporated in the numerical scores they observe, and instead take them at face value: thus,

$$\mathbb{E}^t_{NSL}(Q_i|\mathcal{R}_i^{t-1}, \mathcal{R}_{-i}^{t-1}) = \mathbb{E}(\mathcal{R}_i^{t-1}) - \mathbb{E}(\mathcal{R}_{-i}^{t-1}), \quad \forall\ Q_i, s_i, P_i, Q_{-i}, s_{-i}, P_{-i}.$$

If all consumers were rational, biases in $\mathbb{E}(\mathcal{R}_i^{t-1})$ would be inconsequential. However, the empirical literature (e.g., De Langhe et al. [2015] and Powell et al. [2017]) has often documented the opposite, which justifies the naïve category in here. Setting $\mathbb{E}_{NSL}(Q_i|\mathcal{R}_i^{t-1}, \mathcal{R}_{-i}^{t-1}) = \mathbb{E}(\mathcal{R}_i^{t-1}) - \mathbb{E}(\mathcal{R}_{-i}^{t-1})$ is the most straightforward heuristic.

To sum up:

$$\mathbb{E}^t(Q_i) = \begin{cases} 0 & \text{w/o } \textit{Social Learning} \\ Q_i & \text{w } \textit{Rational Social Learning} \\ \mathbb{E}(\mathcal{R}_i^{t-1}(Q_i, P_i, P_{-i}, s_i, s_{-i})) - \mathbb{E}(\mathcal{R}_{-i}^{t-1}(Q_i, P_i, P_{-i}, s_i, s_{-i})) & \text{w } \textit{Naïve Social Learning} \end{cases}$$

It is useful to define the bias in ratings as the gap between the actual difference in quality

[20] This rrepresents a good approximation for markets – such as the one for books or movies – in which most products accumulate hundreds or thousands of reviews relatively fast.

and the one in ratings, or equivalently, between naïve and rational consumers' expectations:

$$\Theta_i^t := Q_i - \mathbb{E}(\mathcal{R}_i^{t-1}) + \mathbb{E}(\mathcal{R}_{-i}^{t-1}).$$

Denote by $\alpha \in (0,1)$ the share of naïve social learners. We have the following:

Proposition 5 (Reputation is Self-Defeating)**.** *Whenever some consumers are naïve ($\alpha \in (0,1]$), Θ_i^t is decreasing in Θ_i^{t-1}. Moreover,*

$$\frac{\partial^2 \Theta_i^t}{\partial \Theta_i^{t-1} \partial \alpha} < 0.$$

The previous Proposition formalizes the following concept: the better a product's relative reputation in any given period, the worse it will be in the following one. The reason for this is that, whenever a share of consumers are naïve ($\alpha > 0$), an excessively high relative rating for one product will cause disproportionately high sales in the subsequent period. Nevertheless, many of these buyers will not be impressed by the product, because they are only weakly matched with it. Were they to interpret reviews rationally, they would have chosen its alternative. The product's ratings will thus suffer in the subsequent period.

There is one important reason for stressing this implication of my model: traditionally, models of social learning feature a *riches get richer* pattern. This pattern can be extreme: in observational learning (see Banerjee [1992] and Bikhchandani et al. [1992]), for instance, social learning effectively stops after a finite (and often very small) number of periods, due to informational cascades. From then on, one product would capture all consumers. Moreover, with positive probability these cascades select the suboptimal option[21].

In observational learning models, agents learn from the *actions* of their predecessors. Thus, agents have no way to communicate disappointment with their choices. One could assume whether allowing them this possibility solves the problem. However, other existing models of learning from reviews highlights the barriers to entry reviews posit (Vellodi [2018]), as well as the unwillingness of consumers to experiment with unproven options (Kremer et al. [2014], Papanastasiou et al. [2017]). Therefore, they also feature increasing market concentration over time.

In reality, ratings tend to be fairly uniform and uniformly upward-biased, and external Awards often have perverse effects on products' reputation (Kovács and Sharkey [2014]), and

[21]This is true unless sellers can dynamically adjust their prices, as recently shown by Sayedi [2018]. In my model, the self-defeating nature or product ratings over time does not depend on sellers' dynamic reactions to them.

thus likely on future sales. The above Proposition highlights that selling to more consumers effectively equivalent to selling to *worse* (from a match point of view) ones. Therefore, a period's profit comes at the following one's expense. This effect is stronger the more naïve consumers are present: rational consumers' beliefs stabilise around the correct values in period 1.

Proposition 5 describes an inverse proportion between consecutive periods' ratings, and thus suggests an oscillatory pattern for each product's relative reputation over time. It is natural to ask whether relative ratings converge in the long-run, so that biases are exclusive to the first periods. The following Proposition shows that this is not the case, while on the other hand some correction occurs over time.

Proposition 6 (Long-Run Ratings)**.** *All else equal, short-run ratings are biased in favor of the lower quality / more expensive / niche product. Long-run ratings attenuate, but do not erase, the initial biases.*

3.2.2 Learning about Fit

As discussed in Section 3.1.3 and empirically tested for in Section 4, learning about their idiosyncratic match for each product can be as or more valuable for consumers as learning about vertical quality. However, modelling learning about fit is not as straightforward as modelling the learning process for vertical quality. This is because consumers rate products according to their own taste, which (by the standard *iid* taste shocks assumption) is orthogonal to that of their peers.

It is also important to mention that the relationship between learning about the variance of consumer ratings and learning idiosyncratic fit is far from obvious. The variance of ratings, even in the first best (and unlikely) scenario in which it is properly interpreted, only informs consumers about how polarizing a product is, but says nothing about their match value. In particular, when consumers are risk neutral about their taste match (as in virtually all existing models, including the present one), learning about higher order moments of θ_{ij} is, *per se*, pointless.

Nevertheless, it is intuitive (and apparent in the data) that learning in this dimension is both widespread and quantitatively important. One could think of modelling this process in several ways.[22] One tractable possibility is to assume that each period's ratings contain a signal of the product's fit for each future consumer. That is, in presence of social learning,

[22] A more structural and perhaps realistic alternative I have considered, then abandoned due to lack of tractability, is to model a two-step process in which consumers first learn the design of the product given the variance of ratings, for instance by using the heuristic $\mathbb{E}(\mathbb{1}_{s_i=H}|\mathcal{R}_i^t) = 1 \Leftrightarrow Var(\mathcal{R}_i^t) \leq \bar{V}$ for some $\bar{V} > 0$,

$$\mathbb{E}_j^t(\theta_{ij}) = \begin{cases} \theta_{ij} & \text{with probability } \rho_t(\mathcal{R}_i^t) \\ 0 & \text{with probability } 1 - \rho_t(\mathcal{R}_i^t). \end{cases}$$

Clearly, given this formulation it becomes key to specify properties of the probability function $\rho_t(\cdot)$. In particular, notice that this function takes the entire distribution of ratings as input, and not simply its expected value: this is natural since a simple scalar is unlikely to give each consumer information about their own idiosyncrasies.

Also, I have ruled out the possibility that $\rho_t(\mathcal{R}_i^t)$ depends on the particular consumer, whether because of heterogeneous ability to interpret ratings, or because consumers with a certain taste are more likely to find out whether the product is a match. My findings are robust to the case in which, analogously to Section 3.2.1, a fraction α of consumers are naïve, provided that naïvete is modelled as receiving less informative signals, that is, by substituting $\rho_t(\mathcal{R}_i^t)$ with $\beta\rho_t(\mathcal{R}_i^t)$, for $\beta \in (0, 1)$. However, since this extension does not offer any additional insight, I will not devote attention to naïve consumers in this particular context.

To proceed, I make the following assumption.

Assumption 2. *ρ_t solely depends on t, for every history of ratings $\{\mathcal{R}_i^0, \mathcal{R}_i^1, \dots\}$. In particular, ρ_t is increasing in t.*

The second part of this assumption is pretty straightforward, and posits that each generation cannot unlearn what the previous generation knew. The first part is needed for analytical tractability, and posits that a different set of reviews should not influence the informativeness to each consumer about her match value. We have the following:

Proposition 7. *Assume Assumption 2 holds. Then, over time:*

- *Consumers experience, on average, decreased quality, but individually higher utility, due to improved matches.*
- *Consumption segregation increases over time.*[23]

and then try to infer their taste for the product given the percentage ρ_{ij}^t of ratings mentioning features they like, $\theta_{ij} = F^{-1}_{\mathbb{E}(s_i)}(\rho_{ij}^t)$. The qualitative message is unchanged.

[23] In a variation of this model in which each consumer buys several books over time, and learning about taste spills over between products (e.g. enjoying a comedy today leads to buying more comedies tomorrow), individual patterns are opposite to aggregate ones: that is, while consumers diverge from each other over time, each consumer's consumption basket becomes increasingly concentrated.

- *Ratings platform-wide inflate.*

First, because consumers now trade off between quality and fit, and the weight given to fit, ρ^t, is increasing, lower quality products progressively gain those consumers for whom they are a good match at the expense of higher quality options. Second, consumers solve these tradeoffs rationally: by revealed preference, they must be better off. Third, because everyone agrees on quality, but not on taste, choices become more differentiated. Last, since ratings reflect utilities, they inflate with them.

3.3 Managerial Implications

The following are straightforward corollaries of the above findings.

- **A "Love for Large Numbers" is Rational.**[24] In recent influential work, Powell et al. [2017] show experimentally that consumers have a strong preference for products which received many ratings.[25] This is evidence, they argue, for poor statistical reasoning: a product with an average of 3.8 out of 5 is more likely to be great the fewer its ratings (fewer ratings means higher variance, thus greater upside). My findings challenge this view: more ratings effectively correspond to a higher burden of proof. Averaging 3.8 over 5.000 ratings is harder than doing so over 25 fans. That is, for a fixed average rating, expected quality is increasing in the number of ratings. In fact, a heuristic that rewards products for both the average and the number of ratings can outperform one solely based on the average and the variance of ratings. For researchers, this suggests an important relationship between observational learning and learning from reviews in horizontally differentiated markets.

- **Price as a Matching Device.** Higher prices will deter everyone but die-hard fans, thus leading to a more favorable consumer self-selection (though possibly a decrease in present revenue).[26] This represents an anti competitive effect of ratings. Formally, define by P_i^t the price solving the FOCs without social learning, and note that in presence of social learning (and a positive fraction of naïve consumers) the FOCs

[24] See Appendix B for a more in-depth analysis of this point.

[25] Relatedly, for a classic economics example in which consumers naïvely proxy quality with previous market shares, see Caminal and Vives [1996].

[26] This is true even in the case in which ratings negatively reflect prices (Luca and Reshef [2019]). However, the optimal price is, of course, different in that case. See Appendix C for a brief discussion.

contain an additional term given by

$$\frac{\partial \mathbb{E}^N_{t+1}(Q_i|\mathcal{R}^t_i)}{\partial P^t_i} - \frac{\partial \mathbb{E}^N_{t+1}(Q_{-i}|\mathcal{R}^t_{-i})}{\partial P^t_i},$$

where $\mathbb{E}^N_{t+1}(\cdot)$ denote the expectations of $t+1$ generation naïves. This term is positive for every P^t_i, because $\mathbb{E}(\theta_{ij}|j \in \mathcal{J}_i)$ is increasing P^t_i (and $\mathbb{E}(\theta_{-i,j}|j \in \mathcal{J}_{-i})$ is decreasing). Thus, the new optimal price for a patient firm will be larger: $P^t_{i,SL} > P^t_i$ for every i and t.

- **Tendency towards Personalisation.** All else equal, Proposition 1 shows that mainstream products have a hard(er) time creating uniformly positive word of mouth. Therefore, sellers are incentivized to move towards more niche designs. This is particularly true for *i*) low quality, *i*) new and *iii*) more patient sellers. Furthermore, notice that personalisation is associated with an increase in local monopoly power: this provides an additional channel for price increases.

- **Returns to Targeted Advertising Go Up.** By attracting precisely the type of consumers the firm believes to have a higher taste for its products, targeted advertising offers future reputational benefits, on top of current revenue ones.

- **Winner does *Not* Take All.** Typically, dynamic consumer choice models (with or without social learning) feature a positive intertemporal internality for firms. That is, more sales today lead to more sales tomorrow (whether due to switching costs, social influence, informational cascades, consumer inertia, . . .). In my model, future reputation is decreasing in current one, which – at least when a fraction of consumers are naïve – *leads to a less concentrated market*.

- **The Impact of Fake Reviews[27] is Short-Lived.** It follows from the above point that, even assuming that fake reviews are costless to post, and guaranteed to work, their effectiveness might be lower than previously thought. Excessively high ratings will attract consumers who are likely to be disappointed – because of a lack of a strong match with the product. Thus, while biased, long-run ratings are *robust* to fraudulent manipulation.

[27] See Mayzlin et al. [2014] and Luca and Zervas [2016] for two empirical papers documenting the phenomenon.

- **The Problem of Entrants.** Suppose consumers know more about the incumbent (product 0) than the entrant (product 1). Then, they will buy product 1 if $\mathbb{E}(Q_1) + \mathbb{E}(\theta_1) \geq Q_0 + \theta_0$, or equivalently whenever $\theta_0 + Q_0 \leq \mathbb{E}(Q_1)$. Then, average ratings favor the incumbent: $\mathbb{E}(\mathcal{R}_0) = Q_0 + \mathbb{E}(\theta_0 | Q_0 + \theta_0 > \mathbb{E}(Q_1)) > Q_0$ and $\mathbb{E}(\mathcal{R}_1) = Q_1$. The incumbent has a *matching advantage*, which is exacerbated exactly by the introduction of new options, particularly when of high quality.

4 Data

I provide evidence for my theoretical claims using data from the book market. This industry is a perfect fit for the present work. Books display large vertical and horizontal differentiation, and the vast amount of books available on the market makes it necessary for consumers to ease their search process by relying on the opinions of others. Indeed, existing papers (e.g., Chevalier and Mayzlin [2006], Sun [2012]) have demonstrated the importance of consumer ratings in this market.

The book market is also a sizeable one: 674 million print books are sold yearly in the US, and 67% of the population reports having read at least one book in the past year, with pleasure being the main motivation. The average American spends \$110 a year on books, and the book publishing industry has netted a profit of \$26 billion in 2018 (audiobooks adds \$162 million units and \$2.5 billion in profit). There is currently an estimated 45.000 authors in the US.[28] Unsurprisingly, the largest book retailer in the world is Amazon, which lists over 60 million books. The average brick and mortar store in the US stores around 8.000 books; there is large heterogeneity around this number.[29]

To perform my empirical analysis, I collect and match data scraped from two prominent book review platforms, *Goodreads* and Book Marks. *Goodreads* focuses on consumer reviews; *Book Marks* on professional critics'. Last, I complement it with hand collected data on publishers.

This combination provides substantial advantages: first, by comparing *Goodreads* to an external source of information, I do not need to take a stance on – or estimate – books' characteristics. Rather, I will measure the main feature of *Goodreads* ratings against those of Book Marks'. While *Book Marks* is likely not perfectly "objective" – critics are likely to express their personal views, too – it still represents an important benchmark to compare

[28] `https://www.statista.com/topics/1177/book-market/`
[29] `https://www.librarything.com/topic/40375`

Goodreads to. Moreover, differences between the two platforms will be informative of which books are chosen when each is more prominent. Last, *Goodreads* offers extremely granular data about its individual users, and allows me to measure the evolution of their choices and reviews, as well as what kind of information describe most useful.

I briefly describe each of these here. For a more accurate description, see Appendix D.

4.1 Consumer Ratings Data: *Goodreads*

Goodreads launched in 2007 and is by far the world's most popular consumer-generated book reviews platform. In 2019, it averaged over 430 million page views and 50 million unique visitors a month, while boasting over 2.5 billion books.[30] On March 28th, 2013, Amazon announced it had acquired *Goodreads*. The undisclosed fee was rumoured to be around $200 million.[31]

The vast majority of information contained on *Goodreads* is user generated. Upon joining the platform, users can input basic information, such as their gender, age and location, together with a short bio and other information about their reading habits. Books are usually added by either publishers or independent writers to increase their visibility.

Users can rate (with an integer score between 1 and 5), review (with such score plus text), shelf or suggest book to friends. Members can form groups, follow others members, directly search for books they have discovered somewhere else, or browse through the several lists directly curated by *Goodreads*.

Goodreads also provides tailored recommendations to its users, based on their previously read books, and on how they rated them. Although this feature of the website will not be the primary focus of this paper, it is important to notice that – since users arguably value the recommendation role of the website as much as (if not more than) its social aspect – it provides strong incentives for truth telling in rating behavior. This at least alleviates concerns regarding strategic rating behavior, and offers a rationale for my modeling assumptions (I also discuss the robustness of my theoretical results to alternative modelling choices in Appendix C).

[30] See `https://www.goodreads.com/advertisers` and `https://www.goodreads.com/about/us`.

[31] At the time of acquisition, Russ Grandinetti (then Vice President of Kindle content at Amazon) discussed how important this integration would be for Amazon's e-book division: *"Amazon and* Goodreads *share a passion for reinventing reading.* Goodreads *has helped change how we discover and discuss books and, with Kindle, Amazon has helped expand reading around the world. In addition, both Amazon and* Goodreads *have helped thousands of authors reach a wider audience and make a better living at their craft. Together we intend to build many new ways to delight readers and authors alike."*

I scraped detailed *Goodreads* information about more than 3.200 books in the spring of 2019. For each book, I observe all aggregate statistics, such as its publisher and date of publishing, its genre(s), writer, original language, average rating, number of ratings, histogram of ratings, number of text reviews, and so forth. When finding a book they deem interesting, users can mark it "To Read" or add it to a shelf. For each book, I observe the number of users who did so, which I interpret as a proxy for future demand.

Moreover, for each book I collected individual ratings and text reviews. Because collecting all of these is not feasible (e.g., several books have over five million ratings), I collected the first and last 100 ratings in chronological order, on top of the first 30 and last 30 reviews (together with their full text, and whether they specify anything in the "Recommend to" column).

Last, and importantly, for each book I collected detailed information about the top reviews, as upvoted by users. This is informative of which type of reviews appear more useful to users, for different books (e.g., are they positive or negative in valence, do they contain more objective or subjective information, and so forth).

To achieve a heterogenous sample of books that is representative of *Goodreads* at large, I mixed a variety of specific – and very different – *Goodreads* lists. These lists includes the most popular *Goodreads* books of all time, as well as a random sample of new books by genre, books dealing with specific topics (e.g., American politics, climate change) and all books present on *Book Marks* (see below). Overall, while my sample is not perfectly representative of the entire platform, these books display substantial heterogeneity in their characteristics, such as the number and average of ratings, genre, publishers (see below) and all other observables, as shown in Table 1 and Figures 6 and 7.

Parallel to this, to gain further understanding of users' experiences (and trajectories) on *Goodreads*, I scraped data from 1000 users. Because this data represents users who appeared on the *Goodreads*' homepage, which tracks the most recent activity, it is biased towards users who are currently active (and have posted enough reviews, see the next paragraph). Nevertheless, there is a large degree of heterogeneity between these users (Table 2).

For each user, I collected information of the first and last 50 books they have read or rated. For each of these books I collect the aggregate information described above. Whenever a rating has been left, I scraped it, as well as the full text review, if available. User data allows me to answer whether users specialize on a genre or broaden their interests over time; whether users become more or less similar to each other; whether they start purchasing books with more (or higher) ratings; and whether (and for which groups) they become more stringent in their ratings.

4.2 Critics Ratings Data: *Book Marks*

Book Marks is a book critics' review aggregator launched in June 2016 by Literary Hub, a daily literary website that started in 2015 with the mission of "celebrating the literary internet". Literary Hub promoted *Book Marks* as "a Rotten Tomatoes for books".

Book Marks collects critics' reviews from over 70 sources, including newspapers, magazines, and websites, and translates their text into a score: "rave", "positive", "mixed", or "pan". It then averages these scores to come up with the overall one for each book. For the sake of comparison with *Goodreads*, I then mapped these scores into 5, 4, 2 and 1 respectively.[32]

The entirety of *Book Marks* was scraped. For each book featured on the website, I collected all the information presented on its page: the full histogram of ratings, and all text reviews, including their dates, writers and affiliations. In total, this amount to 960 books. Each of these books has been matched to its *Goodreads* information, as described above.

Book Marks' limited size is interesting in itself, as it is informative of which books are under the critics' spotlight, and which others are neglected. In particular: are books reviewed on *Book Marks* more popular than the average *Goodreads* book? Do they have better *Goodreads* ratings? Do they disproportionately belong to certain genres?

4.3 Publishers Data

Last, I obtained information about book publishers. This information was collected from a combination of sources, spanning from literary blogs to the more quantitatively focused *Statista*.[33] I first collected over 100 indie publishers, mostly employing curated lists of recommended ones found on blogs and forums. I then did the same for major ones, defined as those with the highest market shares.

While not exhaustive, these external definitions are consistent with *Goodreads*' information. I find that *i*) books from indie publishers accumulate on average one fifth as many reviews as major ones' (notice, this is despite the fact that my *Goodreads* books selection is,

[32]Several alternative, including $5, 3.66, 2.33, 1$ and $5, 4, 3, 2$ have also been used, without any impact on the results.

[33]See `https://thejohnfox.com/2017/09/30-best-small-indie-literary-publishers/`,`https://www.powells.com/post/lists/24-of-our-favorite-small-presses` and `https://www.statista.com/topics/1177/book-market/`.

if anything, biased in favor of the most successful indie books), and *ii*) indie publishers have substantially fewer bestsellers (as defined by books whose number of ratings lies above the $90th$ percentile in my sample) than mainstream ones.

I then add each publisher only present once in my *Goodreads* dataset – and whose sales are appropriately low, a requirement that only cuts one publisher – to the list of indie ones. Altogether, I categorise around 70% of books as either coming from indie or major publishers, and each category accounts for roughly 35%. The remaining 30% of the books come from middle-sized publishers.

5 Empirical Strategy and Results

5.1 Biases in Ratings

5.1.1 Goodreads Understates Quality Differences

First, I provide a test Proposition 3. That is, my aim is to empirically show that *Goodreads* ratings understate quality differences. That is, for two given products $i_1, i_2 \in \mathcal{I}$ with qualities $Q_{i_1} > Q_{i_2}$, on average,

$$\left|\mathbb{E}(\mathcal{R}^G_{i_1}) - \mathbb{E}(\mathcal{R}^G_{i_2})\right| < Q_{i_1} - Q_{i_2}.$$

However, real qualities are not observed. To get around this problem, I follow the literature (see De Langhe et al. [2015] and discussion therein) in proxying quality with an external - and arguably more objective - source of information: *Book Marks* critics ratings. This methodology relies on the idea that critics *i*) self-select less than consumers in their consumption and reviewing behavior (e.g., critics for several outlets are required to review certain books, independently on their interest for them), and / or *ii*) are more objective in their ratings, that is, they weight Q_i more than θ_{ij} (e.g., a critic knowing horrors are not her favourite genre should be relatively more lenient in horror reviews than her intrinsic enjoyment would dictate). Thus, denoting by $\mathcal{C}$ the set of critics, my empirical test becomes

$$\mathbb{E}_{i_1,i_2\in\mathcal{I},i_1\neq i_2}\left|\mathbb{E}_{j\in\mathcal{J}}(\mathcal{R}^G_{i_1}) - \mathbb{E}_{j\in\mathcal{J}}(\mathcal{R}^G_{i_2})\right| < \mathbb{E}_{i_1,i_2\in\mathcal{I},i_1\neq i_2}\left|\mathbb{E}_{c\in\mathcal{C}}(\mathcal{R}^B_{i_1}) - \mathbb{E}_{c\in\mathcal{C}}(\mathcal{R}^B_{i_2})\right|.$$

Using the universe of *Book Marks* rated books as my sample, I find that the left hand side is around 34% smaller than the right one, that is, *Goodreads* understate, on average, the difference in ratings of two products by roughly 34%[34]. There are two important reasons why, if anything, this figure should understate the result. Both are due to the fact *Book*

[34]Figure 9 displays *Goodreads* ratings as a function of *Book Marks'* ones.

Marks critics ratings are an imperfect proxy for quality. First, at least some book critics are likely to self-select in a similar fashion to consumers (*Book Marks* ratings' sources include a variety of independent magazines and literary blogs, which often have partisan views). Second, because of the relative size of *Goodreads* and *Book Marks*, the comparison is made using books that are deemed interesting, or at least noteworthy, by critics. In fact, *Book Marks* ratings are, on average, higher than *Goodreads* ones: 4.04 vs 3.96. A sample including more low quality books – whose consumers ratings are the most upward biased in my model – would likely considerably strengthen the results.[35]

5.1.2 Goodreads Overrate Polarizing Products

I now turn to testing Proposition 1: consumer ratings relatively advantage more polarizing options. First, key to my analysis in both this and the next Section is a definition of polarizing (or niche) products. Since one of the central theses of the present work is that inferring product design from product ratings is problematic (as further shown in the Section 5.1.3), my definition of "polarizing" must be rooted in products' characteristic, and not in their nature of their *Goodreads* ratings.

I consider several, complementary definitions of polarizing products, including *i*) books taking clearly partisan views on controversial issues (e.g., climate change, American politics), *ii*) books of unusual genres, defined as genres that appear at most 5 times in my sample of 3128 books, *iii*) foreign books. To give some intuition for definition *ii*), consider for instance the genre "mountaineering", which appears three times in my sample: while *prima facie* not a "polarizing" topic, mountaineering books – which are often highly technical in nature – partition the set of consumers into die-hard mountaineering fans with a strong taste for them, and a much larger group, which will likely find them completely uninteresting. Thus, if we were to sample the entire consumer population, a high variance in opinions would be observed.

Important theoretical work (Johnson and Myatt [2006], Bar-Isaac et al. [2012]) has suggested that more polarizing products should be of lower quality: this is because – in line with Clemons et al. [2006] and Sun [2012] – a highly polarizing design increases sales only when competing in the vertical dimension is not possible. Put simply, when convincing everyone is not feasible (low Q_i), a polarizing design (high variance in θ_{ij}) helps convincing at least

[35] If *Goodreads* ratings were uniformly higher, the effect would be somewhat mechanical. In this case, consumers would likely be able to correctly infer quality differences from ratings. For instance, *Uber* customers would likely be skeptical of a driver with an average of 4.7 over 500 rides, given that the majority of drivers are rated above 4.9, despite an absolute difference of only 0.2 out of 5.

someone.

Despite their deep implications for market structure, these theories have received virtually no empirical support. The reason, of course, is that they deal with two unobservables: quality and polarization. Consumer ratings offer a potential way around this issue. Proposition 1, however, offers an important caveat: while worse, polarizing products enjoy more upward-biased ratings.

If we were to trust consumer ratings, do polarizing products actually look worse? Consistent with my theoretical predictions, the answer is negative: looking at the entire sample of *Goodreads* books in my dataset, polarizing ones have, on average, higher ratings: 4.01 vs 3.94 (t=1.95). This remains true after controlling for virtually all of the observable book characteristics.

Unlike on *Goodreads*, do mainstream products properly shine on *Book Marks*? The evidence is mixed. While results are less statistically significant than on *Goodreads*, polarizing books also have higher ratings on *Book Marks* (4.06 vs 4.02, $t = 1.26$). There are two potential explanations for this result: the first hinges on limitations of my data, the second on limitations of my assumptions.

I will start with a brief discussion the former: as I have mentioned before, critics are also self-selecting into reading different books, and potentially ideological in their ratings. For example, several literary magazines employ reviews to "spam" books whose views are in line with theirs. One possible solution around this problem, which will be considered in future work, is to further refine *Book Marks* ratings by only selecting well known, and hopefully relatively non-partisan, outlets, such as major newspapers.[36]

The second limitation is that my definition of polarizing books is, of course, imperfect. For instance, while books of unusual genres are more likely than popular genre ones to have idiosyncrasies that make them polarizing, there are plenty of exceptions to this rule. Again, while an imperfect process, my theoretical results, as well as a wealth of anecdotal evidence presented in Section 5.1.3 – suggest that an *ex-post* definition (that is, one that relies on the observed variance of ratings) is even more problematic. Clearly, more research is needed in this area.

[36] At a broader level, this confirms that proxying actual product characteristics with critics' reviews, while arguably the best practical way to measure biases in consumer ratings, is far from perfect, particularly in markets in which taste plays a fundamental role, such as cultural ones. On the other hand, it also highlights a promising avenue for future research: do consumer ratings offer a more or less partisan view than critics' ones? The empirical answer contained in here, which is admittedly very preliminary, seems to suggest the former.

5.1.3 The Variance of Consumer Ratings Does Not Proxy Polarization

Proposition 4 suggests caution in inferring product design from the variance of product ratings: products that are polarizing *ex-ante* need not be so *ex-post*. A product might simultaneously create disagreement among all of its potential *consumers*, while obtaining consensus among its actual *buyers*. In this Section, I provide empirical support for this claim. I start by running the following regression:

$$Var_i = \alpha + \beta X_i + \gamma Niche_i + \epsilon,$$

where Var_i indicates the variance of *Goodreads* ratings, X_i contains books characteristics, such as year of publication, a dummy for major awards winners and the average and number of ratings, and $Niche_i$ represents my *ex-ante* definitions of niche.

Table 3 contains the results. The first thing to notice is that the coefficient associated to $Niche_i$ is not significantly different from 0, suggesting that niche products might not in fact appear polarizing in terms of their ratings. Of course, one possible concern with this null result is that my definition of niche is imperfect, and fails to capture products that are actually polarizing.

This concern is mitigated by three observations. First, note that the coefficient associated to $NumRatings_i$ is positive and significant.[37] This is despite the fact that $NumRatings_i$ is positively correlated to $AvgRating_i$, and the latter is mechanically, for high enough average ratings (and thus in my data, where ratings average around 4 out of 5), associated to a lower variance in ratings. Bestsellers are purchased by a highly heterogeneous set of buyers: the variance in their ratings reflects this heterogeneity.[38]

Second, to ensure my results are not driven by my own *ad hoc* model selection, I run a LASSO regression (Tibshirani [1996]). LASSO implements model selection algorithmically by imposing a linear penalty term on each regression coefficient. While biasing all coefficient downwards, and thus not providing unbiased estimates for their marginal effects, this selects the most important variables, and its results are thus robust to my initial selection of regressors. While more nuanced than in the previous regression, the results are mostly in agreement. For details, see Appendix E, and in particular Tables 11 and 12.

[37]Furthermore, the raw correlation between the number of ratings and the variance of ratings is positive (Figure 10).

[38]I am currently working on additional scrapes formalizing this result: for each book in my sample, I first obtain a random sample of its *Goodreads* reviewers, and then analyze *i*) how heterogeneous they are in their demographics, and reading habits, and *ii*) how "far" the average book they read is from the book - that is, how strong the "taste mismatch" between them and the book.

Last, my findings are clearly in line with anecdotal evidence. For example, books in the *"biblical fiction"* genre have about half the variance in ratings of those in the *"fiction"* genre (0.49 vs 0.95); bestsellers (as defined by books with more than 200 editions) average a variance of 1.03, compared to 0.89 overall; the results get *stronger* as I raise the threshold to 500, 750 and 1000 editions.[39] It is implausible that these patterns are in line with *ex-ante* polarization.[40] Nevertheless, they are at odds with the existing literature on the topic (Clemons et al. [2006], Sun [2012]), and suggest that marketers and researchers should *not* try to infer a product's design from the variance of its ratings.[41]

Polarizing books often *can* be judged by their cover: readers with a distaste for them will avoid them and their ratings will thus not suffer a similar curse as those of more mainstream products, whose inoffensive design will attract many consumers for whom they are only a weak match. Note that in this sense, first and seconds moments of ratings – as studied in Propositions 1 and 4 – are intertwined: a lower variance is equivalent to a greater upward bias in mean. This is because, at least under the – arguably natural – assumption that consumers are more likely to buy the product the better matched with it they are, a more homogenous buyer pool corresponds to a higher average product-consumer match. Thus, the fact that *a priori* polarizing books do not appear to be so from their ratings offers supporting evidence for the findings in 5.1.2.

5.2 Social Learning

As discussed in Section 3, the relationship between ratings' biases and consumers (mis)learning is far from obvious. To shed light on this issue, I combine two different datasets. First, I leverage user data to directly study *Goodreads*' users behavior over time. Second, I use the fact that *Goodreads* users can directly upvote the most useful reviews for each book to infer what kind of information they deem the most valuable.

[39]The last, extremely selective, group includes books such as "The Jungle Book" (which has an above average variance of 0.95) and "Alice in Wonderland" (1.05).

[40]Furthermore, the top 10 books appearing on *Goodreads*' own curated "Most polarizing" ranking – `https://www.goodreads.com/list/show/6199.The_Most_Polarizing_Books_Of_All_Time` – average over 1.3 million ratings, and none of them are of unusual genre or contain clearly polarizing features like graphic violence of political content.

[41]There are, of course, exceptions, such as products rated by random consumers, perhaps as part of an experimental launch, or more generally products whose polarizing nature is unknown to buyers. However, the latter seems implausible with books, as well as several other (cultural and not) markets.

5.2.1 User Behavior Over Time

In this Section, I study how both reading and rating habits evolve as users spend time on *Goodreads*. It is important to stress that, since participation on *Goodreads* is endogenous (product discovery likely being one of its main motivations), these results need to speak to the platform's causal impact on the average reader. Nevertheless, it is also fair to assume that the alternative technologies for discovering of new and lesser known products (such as reading clubs and internet blogs) would not be nearly as efficient and comprehensive as *Goodreads*, so that *Goodreads*' growth is likely to at least accelerate the formation of a long tail.

For the entirety of the 1001 users scraped, the first 50 (henceforth, Early) and the last 50 (Late) actions (whether a rating, a textual review, or simply information about their current reads) are collected, together with demographics information such as their location, gender, years active and number of ratings left.

First, I analyze how the reading and rating habits of this fixed set of consumer compare in the Early and the Late stage. Figure 13 shows the empirical CDFs of book popularity (as proxied by their number of ratings) for both Early and Late. The distribution in Early clearly first order stochastic dominates that of Late: Early consumers read books that are much more popular. The magnitudes are striking: the average and median number of ratings of books experienced by Late are 81% and 93% lower, respectively.

Moreover, the results are at least partly driven by consumption of bestsellers (books with a number of rating above the $90th$ percentile) and obscure books (number of rating above the $90th$ percentile): I find that, on average, consumers read 235% more obscure books in their last 50 *Goodreads* books compared to their first 50. Conversely, they read 69% fewer bestsellers.[42]

Similarly, Figure 14 shows the empirical CDFs of book quality (as proxied by the average of their ratings) for both Early and Late. If consumers were on *Goodreads* to find the highest quality books, the latter should first order stochastically dominate the former. Instead, the two CDFs follow very similar distributions – and the average quality in Early is actually slightly higher than in Late. This is in line with Proposition 7.

[42]Experienced consumers read, on average, 5.5 obscure books and 5.8 bestsellers. For new consumers, the averages are 1.6 and 18.4. Note that by definition the set of bestsellers and obscure books have, by construction, the same cardinality; thus, one way to interpret the results is that while experienced consumers are almost equally likely to read books in the two categories, new consumers are over ten times more likely to read bestsellers.

Note that my model suggests that a slight decrease in observed ratings when moving from more mainstream to more niche books actually means a larger decrease in experienced quality, given that the latter have more upward-biased ratings.

Because consumers do not seem to simply sort books by their average ratings, they must find other dimensions of *Goodreads* information useful. One possibility is that, while they are prone to experimenting early, they gradually realize their preferences and thus specialize over fewer (and possibly lesser known) genres. To this end, I categorise genres as either popular or obscure, depending on their relative frequency.

I define obscure genres as those that appear at most 200 times out of 100,100[43], and by popular ones the most frequent (e.g., "Fiction", "Fantasy", "Romance", "Non-Fiction", "Mystery", "Young Adult", "Classics"). I find that the number of books in the two categories are 1784 and $38,742$ respectively for Early, compared to 4397 and $24,941$ for Late: upon accumulating *Goodreads* experience, consumers read an increasing share of lesser known genres.

Next, I investigate rating pattens over time. I define by "Leniency" of a rating the difference between the user's rating of each book and the book's average rating. While not perfect – since the average rating, I argue, is biased in the first place – this is useful to proxy consumers' enjoyment of the different books they read. The first, striking fact is that consumers get less lenient in their ratings when they are experienced.

Bondi and Stevens [2018] document, in the context of movies and TV series, that experienced users are more stringent. There, experienced and inexperienced users were two completely disjoint groups. Here, however, the same consumers appear to become stricter over time, suggesting that the results in Bondi and Stevens [2018] are driven by nurture more so than nature.

If *Goodreads* users employ the platform to refine their matches and find lesser known products that are a good fit for them, can I gather any evidence for this improved match over time? One way to disentangle their decreasing leniency over time from their strong taste for specific product is to examine how their leniency changes when they are rating obscure books. To this end, I ran the following users fixed effects regression:

$$Leniency_{ij} = \alpha + \beta \cdot UnusualGenre_i + \gamma Late_j + \delta UnusualGenre_i * Late_j + \epsilon_{ij}.$$

The results are shown in Table 5. Despite being more stringent on average, experienced

[43] There are 153 of these; examples include "Biblical Fiction", "Mountaineering", "Occult", "Apocalyptic".

consumers are actually *less* stringent with books of unusual genre. This suggests that they are indeed able to find products that are a strong match for them.

One important consequence of this individual specialization is collective divergence. Because consumers use information to find what is subjectively *good for them*, rather than what is objectively *good*, their consumption bundles should become more different over time, despite the fact that *Goodreads* does *not* personalize ratings. Whether consumer reviews create convergence or divergence between consumers is an important open question; existing literature on the topic has reached contradicting conclusions (Clemons et al. [2006], Hosanagar et al. [2013]).

To test for this, I compute the average number of books shared by two consumers in the Late and Early group. Upon obtaining the frequency $N_i \geq 1$ for each book, the total number of overlaps is given by

$$\sum_{i \in \mathcal{I}} \binom{N_i}{2} = \sum_{i \in \mathcal{I}} \frac{N_i(N_i - 1)}{2}.$$

I find that consumers in "Early" on average share 6.15 more books than consumers in "Late" (0.959 and 0.156 respectively). Notice, this is despite the fact that activity in "Late" is fairly concentrated over time, as it collects the last 50 actions from users who were active in the first half of 2019. On the other hand, some of these consumers joined in 2007, others in 2018, so that "Early" spans over a decade. This suggests that I am, if anything, underestimating collective divergence.

I test for the robustness of my findings in two different ways. First, it might be that the selection of books of consumers in "Early" and "Late" is driving the results. For example, when measuring leniency, it might be that the baseline average of ratings I am comparing individual ratings to varies systematically across the two groups. To at least partly control for this, I repeat my analysis using only books that are read by users in both groups. The results are robust; see Appendix E.

My second robustness checks take advantage of the fact that "Late" masks substantial heterogeneity. Whether for some consumers the 50 actions collected in Late might span, say, their 51st-100th books read or rated during their second year on *Goodreads*, for others this figure might be the 3051th-3100th books in their tenth year.

If users employ *Goodreads* to specialize and "find their own niche" as the first part of my analysis suggests, then the effects should be stronger for those consumers with more extensive *Goodreads* experience, whether proxied by their number of ratings or by their years of activity. This is exactly what I find, as shown in detail in Appendix E.3.2.[44]

5.2.2 Demand for Information

To offer additional evidence about Goodread's impact on choice, I study the social learning process directly. To this end, I analyze which information is most valuable to consumers, and how this varies across books. I do so by exploiting one particular feature of *Goodreads*, which is that consumers can upvote the text reviews that they find the most helpful.

For each book, a total of 90 reviews were scraped, including their numerical rating and full text. The 90 reviews are equally distributed between "Oldest" (the first 30 reviews posted about each book, in chronological order), "Newest" (last 30) and "Default". The 30 "Default" reviews are the ones that were most upvoted by consumers, and are thus prominently displayed on each book's *Goodreads* page.

While a full-fledged text analysis is beyond the scope of this paper, the data allows me to categorize reviews beyond their numerical score. On top of the text, each review has a "Recommended To" entry in which users can, in a few words, indicate a set of their peers who they think would enjoy the book. This entry can range from one size fits all recommendations (e.g., "everyone!", "nobody", "you") to more tailored ones (e.g., "comedy lovers", "fans of Dan Brown", "those who miss California"). While the former provide information about a book's quality, the latter serves a matching purpose, simultaneously providing good news for certain users, and negative ones for others.

I ask two set of questions. First, how does the numerical information contained in text reviews ("Default" and otherwise) differ from that contained in all numerical ratings? Second, what determines a review's usefulness? Does the answer vary with books characteristics?

To answer the first, for each book, I compare the average numerical rating, as well as the variance of ratings, of the different review groups to each other as well as to the aggregate one for the book. I test this using the simple specification:

$$NumStars_{ij} = \alpha + \beta_1 Oldest_{ij} + \beta_2 Default_{ij} + \gamma_1 NumReviews_i + \gamma_2 AvgRating_i + \epsilon_{i,j},$$

where $NumStars_{ij}$, $Oldest_{ij}$ and $Default_{ij}$ indicate the numerical rating, timing and popularity of each of the 90 reviews and $NumReviews_i$ and $AvgRating_i$ refer to book char-

[44] One part of the analysis in progress looks at the "Early" behavior of users as a function of their future number of ratings, to understand whether they were different *before* joining the platform. While informative, this method still has limitations: in particular, one source of endogeneity my data would not be able to tease out would be given by users who have similar taste and knowledge when joining *Goodreads*, but different levels of curiosity and drive to discover new products.

acteristics. I find that reviews are, on average, less lenient than numerical ratings: users seem to be more motivated to take time to write negative information (Table 6).

This can be interpreted in two ways: the first is that consumers feel more strongly about books they did not like. The second is that, because most books have a high average rating, consumer can only justify spending the amount of time needed to write a review whenever this moves the posterior of future readers enough – that is, when it is negative. Notice how the latter form of self-selection favour lower quality products, and thus, while absent from my theoretical model, would corroborate its findings.[45]

Because reviews contain more negative information, they also have a higher variance. The exception to both of these facts is represented by the oldest reviews. This should not come as a surprise. The first 30 reviewers for each book are extremely self-selected, and usually contain a mix of die-hard fans (Li and Hitt [2008]) and fraudulent raters (Luca and Zervas [2016]).

Turning to the second – and more substantial – point, I find that reviews that are deemed more useful by consumers also have precise features. As shown in Table 7, some of them are intuitive: controlling for a book's overall popularity as expressed by the number of reviews it has received[46], they are on average longer, more negative[47], and more likely to contain a "Recommended To" section.

Last, I analyze the importance of tailored recommendations – which are more informative about a book's match value for different consumers than they are about its quality – and whether it depends on the book's type. To this end, I create three dummies, "Recommended for Everyone", "Recommended for No One" and "Recommended for Some". The latter is the subset of "Recommended for" that specifies a non-trivial set of consumers. This dummy is created manually by matching each "Recommended for" entry to words that are symptomatic of matching information, as found on a subset of ratings. Examples include "those who", "interested", "if you liked", "fans", "want", "seekers".

I then run two separate regressions aimed to determine how the presence of match-relevant information ("Recommended for Some") influence reviews usefulness for both bestsellers and unknown books:

$$\begin{aligned} NumLikes_{ij} = \alpha + \beta_1 RecommendedSome_{ij} + \beta_2 NumStars_{ij} \\ + \beta_3 CharLength_{ij} + \gamma_1 NumReviews_i + \gamma_2 AvgRating_i + \epsilon_{i,j}. \end{aligned}$$

[45] See Appendix C.1.2 for a brief discussion of motivated ratings.

[46] The results do not change when using $NumRatings_i$ or $ToRead_i$.

[47] Again, more negative information could be more relevant either due to an intrinsic negativity bias or because the average ratings are mostly high.

Tables 8 and 9 contain the results. While a specific recommendation ("Recommended for Some") predicts usefulness for unknown books' reviews, this is not the case for bestsellers'. One way to interpret these results is that with unknown books, consumers prioritize learning about match quality, while the opposite is true with bestsellers, for which learning is mostly about vertical quality – either because match quality is known, or because learning about match quality is deemed less important for these books to begin with (because their success suggests the average consumer is likely to enjoy them).

Another substantive finding is that the while a high numerical rating predicts reviews usefulness for unknown books, the opposite is true for bestsellers. This is in line with readers learning from *bad news* in the case of bestsellers (e.g., by visiting the page with a prior of buying, and changing their mind in the face of negative information), and from *good news* for unknown books (where good news could be either about vertical or horizontal attributes).

These findings corroborate those in 5.2.1. Taken together, they suggest that *Goodreads*' ratings usefulness to consumers lies more in its *matching* function than in its *sorting* one. This is particularly true for lesser known products, for two potential reasons. First, lesser known products are also more polarizing[48], which increases the relative importance of horizontal information. Second, while consumers might be uncertain about the quality of each book, they might have a more precise sense of the horizontal features of better known ones (e.g., they received press coverage, or their author specializes in a given genre).

6 Conclusions and Future Research

This paper studies social learning from consumer reviews in a vertically and horizontally differentiated market, both theoretically and empirically, using book ratings data. *A priori*, social learning can operate through two distinct (and sometimes opposing) channels: on one hand, it can help consumers *sort* products according to quality, so to yield them *objectively better* products; on the other hand, it can *match* consumers with products they have a strong taste for, so to yield them *subjectively better* matches.

First, I document the very different structure of (and systematic biases in) aggregate ratings in these two scenarios. When consumers screen products mostly based on their vertical quality, a natural bias emerges in that the highest quality products' average ratings are lowered by the high number of weakly matched consumers these products attract ("the

[48]E.g., they are more likely to be of unusual genre, foreign, or produced by indie publishes, who often focus on very specific subgenres.

curse of the bestseller"). For example, a consumer does not need to be a comedy fan to read a comedy advertised as the best one ever written. Still, she will probably not enjoy it as much as she will an excellent book of her favorite genre.

When, instead, consumers use reviews to match with products based on fit, I show that consumer ratings advantage products that are more polarizing: these products are either loved or hated by the average consumer, so they are disproportionately likely to be loved by their average *buyer* (and thus *reviewer*). Products with a mainstream design fail to attract an equally self-selected group of buyers. Relatedly, I show that the variance of ratings contains very little information about the *ex-ante* nature of each product – a finding the reverses both the existing literature (Clemons et al. [2006], Sun [2012]) and the way platforms themselves interpret this information[49], and which inform researchers and managers alike.

Empirically, using data from *Goodreads*, I find that the platform serves much more of a matching purpose than a sorting one. *Goodreads* users read and rate increasingly obscure books of unusual genre, often compromising on their average ratings, as they get more experienced. Interestingly, some of the reviewers are internalising this. Kelly Jensen, a writer and *Goodreads* user, argues on her blog that the five star system is obsolete: "I might not be the reader who needs a specific book, but I know there's a reader out there who does".[50] Pop culture and sports website *The Ringer* acknowledges a similar difficulty in ranking TV series.[51] Indeed, *Netflix* has long moved away from simply ranking its most popular movies and TV series, and has instead built its business model around incredibly specific machine learning driven personalized recommendations.[52]

I see multiple avenues for future research, both theoretical (as discussed in Appendix C) and not. In particular, more attention should be devoted to the distributional consequences of consumer ratings platforms, an important topic on which the existing literature has not yet reached an agreement. The present work suggests that – even absent explicit personalization (e.g., adaptive numerical ratings) – in contexts in which product fit is key platforms fragment the market, favoring lower quality and more polarizing products, and creating small and highly specialized groups of consumers who look increasingly unlike each other. As discussed throughout the paper, this carries fundamental implications for writers, platforms and retailers alike.

[49] https://www.goodreads.com/list/show/6199.The_Most_Polarizing_Books_Of_All_Time
[50] https://bookriot.com/2018/07/09/star-ratings-on-goodreads/
[51] https://www.theringer.com/tv/2018/12/3/18123279/2018-year-in-television
[52] See for instance https://www.latimes.com/business/hollywood/la-fi-ct-tca-netflix-cindy-holland-20180729-story.html.

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.

Heski Bar-Isaac, Guillermo Caruana, and Vicente Cuñat. Search, design, and market structure. *American Economic Review*, 102(2):1140–60, 2012.

Ron Berman and Zsolt Katona. Curation algorithms and filter bubbles in social networks. 2018.

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.

Tommaso Bondi and Ryan Louis Stevens. The good, the bad and the picky: Consumer heterogeneity and the reversal of movie ratings. 2018.

Leif Brandes, David Godes, and Dina Mayzlin. Extremity bias in online reviews: a field experiment. 2018.

Brightlocal. Local consumer review survey. 2018.

John W Byers, Michael Mitzenmacher, and Georgios Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 248–265. ACM, 2012.

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

Luis Cabral and Gabriel Natividad. Box-office demand: The importance of being# 1. *The Journal of Industrial Economics*, 64(2):277–294, 2016.

Ramon Caminal and Xavier Vives. Why market shares matter: An information-based theory. *The RAND Journal of Economics*, pages 221–239, 1996.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2017.

Pei-Yu Chen, Yili Hong, and Ying Liu. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science*, 64(10):4629–4647, 2017.

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.

Pradeep K Chintagunta, Shyam Gopinath, and Sriram Venkataraman. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957, 2010.

Eric K Clemons, Guodong Gordon Gao, and Lorin M Hitt. When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of management information systems*, 23(2):149–171, 2006.

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*, 42(6):817–833, 2015.

Chrysanthos Dellarocas and Ritu Narayan. Tall heads vs. long tails: Do consumer reviews increase the informational inequality between hit and niche products? *Robert H. Smith School of Business Research Paper*, (06-056), 2007.

Loretti I Dobrescu, Michael Luca, and Alberto Motta. What makes a critic tick? connected authors and the determinants of book reviews. *Journal of Economic Behavior & Organization*, 96:85–103, 2013.

Apostolos Filippas, John Horton, and Joseph Golden. Reputation inflation: Evidence from an online labor market. *Work. Pap., NYU*, 1, 2018.

Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, 90(2): 217–232, 2014.

Duncan Sheppard Gilchrist and Emily Glassberg Sands. Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5):1339–1382, 2016.

David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4):545–560, 2004.

Rick Harbaugh, John Maxwell, and Kelly Shue. Consistent good news and inconsistent bad news. *Wokring Paper*, 2016.

Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, 2013.

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

Justin P Johnson and David P Myatt. On the simple economics of advertising, marketing, and product design. *American Economic Review*, 96(3):756–784, 2006.

Balázs Kovács and Amanda J Sharkey. The paradox of publicity: How awards can negatively affect the evaluation of quality. *Administrative science quarterly*, 59(1):1–33, 2014.

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. 2016.

Michael Luca and Oren Reshef. What are consumers rating? evidence from *Yelp!* 2019.

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

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

Enrico Moretti. Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies*, 78(1):356–393, 2011.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

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, 2017.

Tiberiu Popoviciu. Sur l'approximation des fonctions convexes d'ordre supérieur. *Mathematica (Cluj)*, 10:49–54, 1935.

Derek Powell, Jingqi Yu, Melissa DeWolf, and Keith J Holyoak. The love of large numbers: a popularity bias in consumer choice. *Psychological science*, 28(10):1432–1442, 2017.

Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental economics*, 9(2):79–101, 2006.

Amin Sayedi. Pricing in a duopoly with observational learning. *Available at SSRN 3131561*, 2018.

Lones Smith and Peter Sørensen. Pathological outcomes of observational learning. *Econometrica*, 68(2):371–398, 2000.

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

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Catherine Tucker, Juanjuan Zhang, and Ting Zhu. Days on market and home sales. *RAND Journal of Economics*, pages 337–360, 2013.

Raluca M Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

Stefano Vaccari, Costis Maglaras, and Marco Scarsini. Social learning from online reviews with product choice. 2018.

Nikhil Vellodi. Ratings design and barriers to entry. *Available at SSRN 3267061*, 2018.

Jared Watson, Anastasiya Pocheptsova Ghosh, and Michael Trusov. Swayed by the numbers: the consequences of displaying product review attributes. *Journal of Marketing*, 82(6): 109–131, 2018.

Fang Wu and Bernardo A Huberman. How public opinion forms. In *International Workshop on Internet and Network Economics*, pages 334–341. Springer, 2008.

Ya You, Gautham G Vadakkepatt, and Amit M Joshi. A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, 79(2):19–39, 2015.

Juanjuan Zhang. The sound of silence: Observational learning in the us kidney market. *Marketing Science*, 29(2):315–335, 2010.

A Proofs

Proof of Proposition 1

To prove this result, I need a basic technical Assumption:

Assumption 3. *The function $\theta F_L(\theta)$ is weakly convex.*

This Assumption is satisfied by most well known distributions, including the uniform, the normal (and all discrete distributions).

To prove the desired inequality, I start by defining the two conditional density functions,

$$f_L^{max}(\theta) := f_L(\theta|\theta_L > \theta_H)$$

and symmetrically for $f_H^{max}(\theta)$.

By independence, we have that

$$f_L^{max}(\theta) = \frac{f_L(\theta) \cdot Prob(\theta_H \leq \theta)}{P(\theta_L > \theta_H)} = \frac{f_L(\theta) \cdot F_H(\theta)}{M_L}.$$

Therefore,

$$E(\theta_L|\theta_L > \theta_H) = \frac{1}{M_L}\int_{\underline{\theta}}^{\bar{\theta}} \theta f_L(\theta) F_H(\theta) d\theta = \frac{1}{M_L}\int_{\underline{\theta}}^{\bar{\theta}} \big(\theta F_H(\theta)\big) f_L(\theta) d\theta = \frac{1}{M_L} \cdot \mathbb{E}_L\big(\theta F_H(\theta)\big),$$

where $\mathbb{E}_L$ indicates that the expectation is taken over the $F_L(\cdot)$ distribution. Symmetrically, we have that $E(\theta_H|\theta_H > \theta_L) = \frac{1}{M_H} \cdot \mathbb{E}_H(\theta F_L(\theta))$.

Given Assumption 1, we are left with having to prove that

$$\mathbb{E}_L\big(\theta F_H(\theta)\big) > \mathbb{E}_H\big(\theta F_L(\theta)\big).$$

Proving the inequality directly is hard given that we are taking expectations of two different functions under two different distributions. Therefore, instead of comparing the two directly, we compare each of them to $\mathbb{E}_L\big(\theta F_L(\theta)\big)$, and try to show that

$$\mathbb{E}_L\big(\theta F_H(\theta)\big) > \mathbb{E}_L\big(\theta F_L(\theta)\big) \geq \mathbb{E}_H\big(\theta F_L(\theta)\big).$$

Under Assumption 3, the second inequality follows from the fact $F_H(\cdot)$ SOSD $F_L(\cdot)$. This is because for any two random variables such that the first SOSD the second, expectations of convex functions are larger under the second. To prove the first, appeal to the definition of rotation directly: $F_L(\theta) > (<) F_H(\theta)$ for every $\theta < (>) \theta^\dagger$. In other words, both $\mathbb{E}_L\big(\theta F_H(\theta)\big)$ and $\mathbb{E}_L\big(\theta F_L(\theta)\big)$ are weighted averages of θ under $F_L(\cdot)$ but the second puts more weight on lower values of θ and less on higher ones.

Last, I want to show that $Q_1 > Q_2$ does not imply $\mathbb{E}(\mathcal{R}_1) > \mathbb{E}(\mathcal{R}_2)$, and $Var(\theta_2) > Var(\theta_1)$ does not imply $Var(\mathcal{R}_2) > Var(\mathcal{R}_1)$. To this end, I present a case which simultaneously offers a counterexample to both. Let $Q_1 = 1$, $Q_2 = 0$,

$$\theta_1 = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}$$

and

$$\theta_2 = \begin{cases} -2 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{2}. \end{cases}$$

Notice this example says the two products have the distribution of shocks, with 50% of the population liking them, and the remaining 50% disliking them. However, in both cases, opinions about product 2 are stronger, that is, product 2 is more polarizing.

Therefore, we have $\mathcal{J}_2 = \{j \in \mathcal{J} \mid \theta_{2,j} = 2\}$, and consequently $\mathcal{J}_1 = \{j \in \mathcal{J} \mid \theta_{2,j} = -2\}$. Note that $\mathcal{J}_1$ does not depend on a taste for product 1, but rather on a distaste for product 2. Thus, by independence of taste shocks, we have $\mathbb{E}(\theta_{1,j} \mid j \in \mathcal{J}_1) = 0$. As a result,

$$\mathbb{E}(\mathcal{R}_1) = Q_1 + \mathbb{E}(\theta_{1,j} \mid j \in \mathcal{J}_1) = 1 < 2 = Q_2 + \mathbb{E}(\theta_{2,j} \mid j \in \mathcal{J}_2) = \mathbb{E}(\mathcal{R}_2),$$

$$Var(\mathcal{R}_1) = Var(\theta_{1,j} \mid j \in \mathcal{J}_1) = \frac{1}{4} > 0 = Var(\theta_{2,j} \mid j \in \mathcal{J}_2) = Var(\mathcal{R}_2).$$

This provides a counterexample to both the average and the variance of ratings reflecting *ex-ante* relationships between products. Notice how this result does not depend on the market shares being symmetric: any distribution of the form

$$\theta_2 = \begin{cases} -1 - y & \text{with probability } \rho \\ \frac{\rho(1+y)}{(1-\rho)} & \text{with probability } 1 - \rho, \end{cases}$$

with $\rho \in (0,1)$ and $y > \frac{1-\rho}{\rho} - 1$, would work. ■

Proof of Corollary 1

Proof. We have that

$$\mathbb{E}(\theta_L|\theta_L > c) = \frac{\int_c^{\bar{\theta}} \big(1 - F_L(\theta)\big) d\theta}{1 - F_L(c)}$$

and similarly

$$\mathbb{E}(\theta_H|\theta_H > c) = \frac{\int_c^{\bar{\theta}} \big(1 - F_H(\theta)\big) d\theta}{1 - F_H(c)}.$$

Note that $F_L(\theta^\dagger) = F_H(\theta^\dagger)$ by definition of $\theta^\dagger$, and that $F_L(\theta) < F_H(\theta)$ for every $\theta > \theta^\dagger$. As a result,

$$1 - \int_c^{\bar{\theta}} \big(1 - F_H(\theta)\big) d\theta < 1 - \int_c^{\bar{\theta}} \big(1 - F_L(\theta)\big) d\theta$$

and hence

$$\mathbb{E}(\theta_L|\theta_L > \theta^\dagger) > \mathbb{E}(\theta_H|\theta_H > \theta^\dagger).$$

The result is true *a fortiori* for every $c > \theta^\dagger$. By continuity of both conditional expected values – which follows from the smoothness of both $F_L(\cdot)$ and $F_H(\cdot)$ – we have that there exists a $c^* \in [\underline{\theta}, \theta^\dagger)$ such that the result is true for every $c > c^*$. ■

Corollary 2. *When the market shares are asymmetric ($M_H > 1/2$), the above result extends to* relative *conditional tastes:*

$$\mathbb{E}(\theta_L - \theta_H|\theta_L > \theta_H) > \mathbb{E}(\theta_H - \theta_L|\theta_H > \theta_L).$$

Proof of Corollary 2

First notice that

$$\mathbb{E}(\theta_H) = M_H \cdot \mathbb{E}(\theta_H|\theta_H > \theta_L) + M_L \cdot \mathbb{E}(\theta_H|\theta_L > \theta_H)$$

$$\mathbb{E}(\theta_L) = M_L \cdot \mathbb{E}(\theta_L|\theta_L > \theta_H) + M_H \cdot \mathbb{E}(\theta_L|\theta_L > \theta_H)$$

Thus,

$$\begin{aligned}\mathbb{E}(\theta_H) = \mathbb{E}(\theta_L) = 0 \Rightarrow M_H \cdot \mathbb{E}(\theta_H|\theta_H > \theta_L) + M_L \cdot \mathbb{E}(\theta_H|\theta_L > \theta_H)\\ - M_L \cdot \mathbb{E}(\theta_L|\theta_L > \theta_H) - M_H \cdot \mathbb{E}(\theta_L|\theta_H > \theta_L) = 0.\end{aligned}$$

Rearranging, we get

$$M_H\mathbb{E}(\theta_H - \theta_L|\theta_H > \theta_L) = M_L\mathbb{E}(\theta_L - \theta_H|\theta_L > \theta_H)$$

or, equivalently,

$$\frac{\mathbb{E}(\theta_L - \theta_H | \theta_L > \theta_H)}{\mathbb{E}(\theta_H - \theta_L | \theta_H > \theta_L)} = \frac{M_H}{M_L} > 1,$$

as required. ■

Proof of Proposition 2

The Proof is straightforward. To see that ratings are not inflated, notice that

$$\mathbb{E}(\theta_i | Q_i - P_i > Q_{-i} - P_{-i}) = \mathbb{E}(\theta_i) = 0.$$

Thus,

$$\mathbb{E}(\mathcal{R}_i) = Q_i + \mathbb{E}(\theta_i | Q_i - P_i > Q_{-i} - P_{-i}) = Q_i.$$

To see that the variance in ratings reflects the ex-ante one, notice that

$$Prob(\theta_i \leq c | Q_i - P_i > Q_{-i} - P_{-i}) = Prob(\theta_i \leq c).$$

Which means $F_{s_i}^{\mathcal{J}_i}(\cdot) = F_{s_i}(\cdot)$ and thus $Var_{F_{s_i}}^{\mathcal{J}_i}(\theta_i) = Var_{F_{s_i}}(\theta_i)$. ■

Proof of Proposition 3

Since ratings for a product reflect its quality plus the average match with its buyers, and the two products have the same design, we have

$$\begin{aligned}
R(\tilde{q}_1) &= \tilde{q}_1 + E(\theta_1 | \theta_1 \geq \tilde{q}_2 - \tilde{q}_1 + \theta_2) \\
&< \tilde{q}_1 + E(\theta_1 | \theta_1 \geq \tilde{q}_1 - \tilde{q}_2 + \theta_2) \\
&= \tilde{q}_1 + E(\theta_2 | \theta_2 \geq \tilde{q}_1 - \tilde{q}_2 + \theta_1) \\
&= (\tilde{q}_1 - \tilde{q}_2) + \tilde{q}_2 + E(\theta_2 | \theta_1 \geq \tilde{q}_1 - \tilde{q}_2 + \theta_1) \\
&= \tilde{q}_1 - \tilde{q}_2 + R(\tilde{q}_2),
\end{aligned}$$

where I have used that the conditional expected value is increasing in the lower bound of integration for the inequality in the second line, and symmetry (given θ_1, θ_2 follow the same distribution) in the third. ■

Proof of Proposition 4

I want to show that

$$Var\big(\theta_L|\theta_L > \theta_H + \Delta(Q)\big) > Var\big(\theta_H|\theta_H > \theta_H - \Delta(Q)\big), \quad \forall\Delta(Q) > \Delta^*(Q).$$

First notice that, when $\Delta(Q)$ approaches $\bar{\theta} - \underline{\theta}$, we have

$$Var\big(\theta_H|\theta_H > \theta_H - \Delta(Q)\big) \to Var(\theta_H).$$

On the other hand, the fact that $\theta_L > \theta_H + \Delta(Q)$ implies $\theta_L \in (\underline{\theta} + \Delta(Q),\ \bar{\theta})$. Therefore, Popoviciu's Inequality (Popoviciu [1935]) implies that

$$Var\big(\theta_L|\theta_L > \theta_H + \Delta(Q)\big) \leq \frac{1}{4}\big(\bar{\theta} - \underline{\theta} - \Delta(Q)\big)^2.$$

Notice that the right hand side gets arbitrarily small as $\Delta(Q) \to \bar{\theta} - \underline{\theta}$, implying the existence of a $\Delta^*(Q)$ such that $Var\big(\theta_L|\theta_L > \theta_H + \Delta(Q)\big) < Var(\theta_H)$ for every $\Delta(Q) > \Delta^*(Q)$. ■

Proof of Proposition 5

Proof. To show that Θ_i^t is decreasing in Θ_i^{t-1}, first observe that since Q_i is fixed, the claim is equivalent to $R_i^t - R_{-i}^t$ is decreasing in $R_i^{t-1} - R_{-i}^{t-1}$. The difference $R_i^t - R_{-i}^t$ reflects how well matched product i is compared to its alternative. Because naïve and rational consumers' choices differ when $\theta_i^{t-1} \neq 0$, the average match will reflect the weighted average for the two categories:

$$\begin{aligned}
R_i^t - R_{-i}^t &= Q_i + \alpha E\big(\theta_i \mid \theta_i + R_i^{t-1} - P_i \geq \theta_{-i} + R_{-i}^{t-1} - P_{-i}\big) \\
&\quad + (1-\alpha)E\big(\theta_i \mid \theta_i + Q_i - P_i \geq \theta_{-i} - P_{-i}\big) \\
&\quad - \alpha E\big(\theta_{-i} \mid \theta_{-i} + R_{-i}^{t-1} - P_{-i} \geq \theta_i + R_i^{t-1} - P_i\big) \\
&\quad - (1-\alpha)E\big(\theta_{-i} \mid \theta_{-i} + Q_{-i} - P_{-i} \geq \theta_i - P_i\big) \\
&= Q_i + \alpha E\big(\theta_i \mid \theta_i \geq \theta_{-i} - (R_i^{t-1} - R_{-i}^{t-1}) + P_i - P_{-i}\big) \\
&\quad + (1-\alpha)E\big(\theta_i \mid \theta_i + Q_i - P_i \geq \theta_{-i} - P_{-i}\big) \\
&\quad - \alpha E\big(\theta_{-i} \mid \theta_{-i} \geq \theta_i + (R_i^{t-1} - R_{-i}^{t-1}) + P_{-i} - P_i\big) \\
&\quad - (1-\alpha)E\big(\theta_{-i} \mid \theta_{-i} + Q_{-i} - P_{-i} \geq \theta_i - P_i\big)
\end{aligned} \tag{1}$$

Clearly, Q_i, $(1-\alpha)E\big(\theta_i \mid \theta_i + Q_i - P_i \geq \theta_{-i} - P_{-i}\big)$ and $(1-\alpha)E\big(\theta_{-i} \mid \theta_{-i} + Q_{-i} - P_{-i} \geq \theta_i - P_i\big)$ are independent of $R_i^{t-1} - R_{-i}^{t-1}$. On the other hand, $\alpha E\big(\theta_i \mid \theta_i \geq \theta_{-i} + (R_{-i}^{t-1} - R_i^{t-1}) + P_i - P_{-i}\big)$ and $-(1-\alpha)E\big(\theta_{-i} \mid \theta_{-i} + Q_{-i} - P_{-i} \geq \theta_i - P_i\big)$ are both decreasing in it. This follows from the fact that the conditional expected values of θ_i and θ_{-i} are monotonically increasing in the lower bound of integration.

Showing that

$$\frac{\partial^2 \Theta_i^t}{\partial \Theta_i^{t-1} \partial \alpha} < 0$$

is immediate given that each term is linear and increasing in α.

■

Proof of Proposition 6

Proof. To show that period 1 ratings are biased, appeal directly to Propositions 1 and 3.

I now show that long-run ratings are also biased. I do this by showing that if the ratings two products with either i) different vertical attributes (quality or price) or ii) different designs are unbiased in a given period, they will immediately start displaying biases again in the following one.

In light of Proposition 5, this amounts to showing that 0 is not a fixed point of the mapping $\Theta_i^t(\Theta_i^{t-1})$. Suppose $\Theta_i^{t-1} = 0$. Since Θ_i measures differences in beliefs between naïve and rational consumers, this means that $t-1$ generation naïve and rational consumers fully agree on quality differences.

First, consider two products such that $s_i = s_{-i}$, but (without loss of generality) $Q_i - P_i > -P_{-i}$. Then, because α cancels out and $R_i^{t-1} - R_{-i}^{t-1} = 0$, equation 1 simplifies to

$$\begin{aligned} R_i^t - R_{-i}^t = Q_i &+ E\big(\theta_i \mid \theta_i \geq \theta_{-i} - Q_i + P_i - P_{-i}\big) \\ &- E\big(\theta_{-i} \mid \theta_{-i} \geq \theta_i + Q_i - P_i + P_{-i}\big). \end{aligned}$$

Notice that, since $Q_i - P_i + P_{-i} > 0$, and $s_i = s_{-i}$, $E\big(\theta_i \mid \theta_i \geq \theta_{-i} - Q_i + P_i - P_{-i}\big) - E\big(\theta_{-i} \mid \theta_{-i} \geq \theta_i + Q_i - P_i + P_{-i}\big) < 0$. But this means

$$\Theta_i^t = Q_i - R_i^t + R_{-i}^t > 0.$$

Thus, $\Theta_i = 0$ is not a fixed point in this case.

Next, consider two products such that $Q_i - P_i = -P_{-i}$ and (without loss of generality) $s_i = H \neq L = s_{-i}$. Then, equation 1 simplifies to

$$R_i^t - R_{-i}^t = Q_i + E\big(\theta_i \mid \theta_i \geq \theta_{-i}\big) - E\big(\theta_{-i} \mid \theta_{-i}\big).$$

Proposition 1 showed that the right hand side is smaller than Q_i, again implying $\Theta_i^t > 0$.

Last, I want to show that, in both of these cases, long-run biases have the same direction, but smaller magnitude, than short-run ones.

Notice that in both cases $\Theta_1 < 0$ (this follows from Propositions 1 and 3). Thus, to conclude the proof it only remains to be shown that the fixed point, which I will denote by Θ_∞, satisfies $\Theta_1 < \Theta_\infty < 0$. To show that Θ_∞ is negative, it is sufficient to combine two of the previous findings: first, Θ_t is decreasing in Θ_{t-1}, and second, $\Theta_t < 0$ when $\Theta_{t-1} = 0$.

To show that $\Theta_\infty > \Theta_1$, by negative monotonicity it is enough to show that $\Theta_2 > \Theta_1$. To this end, notice that some naïve consumers will overreact to Θ_1 by purchasing the niche / lower quality product even when suboptimal for them. This will push its rating down, to the advantage of product 2, which experienced excessively low market share. Thus, $\Theta_2 > \Theta_1$, which concludes the proof.

■

Proof of Proposition 7

Proof. I start by showing that the experienced quality experienced by users goes down. For the sake of exposition, take two products with symmetric designs $s_1 = s_2$ and prices, $P_1 = P_2$ (the proof for the case $s_1 \neq s_2$ and $P_1 \neq P_2$ is analogous). Without loss of generality, let $Q_1 > Q_2$. Then,

$$\begin{aligned} M_1^t &= P(Q_1 + \rho^t \theta_1 \geq Q_2 + \rho^t \theta_2) \\ &= P(\rho^t(\theta_{s_1} - \theta_{s-2}) \geq Q_1 - Q_2) \end{aligned}$$

The variable $\xi_1 = \theta_1 - \theta_2$ clearly has mean 0. Therefore,

$$\begin{aligned} P(\rho^t\big(\theta_1 - \theta_2\big) \geq Q_1 - Q_2\big) &= P\Big(\xi_1 \geq \frac{Q_2 - Q_1}{\rho^t}\Big) \\ &= 1 - F_{\xi_1}\Big(\frac{Q_2 - Q_1}{\rho^t}\Big), \end{aligned}$$

which is clearly a decreasing function of ρ^t.

Next, I prove that individual consumers experience (weakly) higher utility over time. The intuition is straightforward: more information can only improve the accuracy of the tradeoffs performed by consumers. To see this more formally, notice that we only need to look at those consumer whose choices are different between period t and $t+1$. In particular, this set can be reduced to the intersection of $\mathcal{J}_1^t \cap \mathcal{J}_2^{t+1}$, since the set $\mathcal{J}_2^t \cap \mathcal{J}_1^{t+1}$ is empty. That is,

$$\mathcal{J}_1^t \cap \mathcal{J}_2^{t+1} = \{j \in \mathcal{J} \mid Q_1 + \rho^t \theta_{1,j} \geq Q_2 + \rho^t \theta_{2,j},\ Q_1 + \rho^{t+1}\theta_{1,j} \leq Q_2 + \rho^{t+1}\theta_{2,j}\}.$$

Clearly, for each consumer in this group, $\rho^{t+1}(\theta_{2,j} - \theta_{1,j}) \geq Q_1 - Q_2 \geq \rho^t(\theta_{2,j} - \theta_{1,j})$, which implies $\theta_{2,j} \geq \theta_{1,j}$. But then *a fortiori*

$$\begin{aligned}
U_j^{t+1} &= Q_2 + \theta_{2,j} \\
&= Q_2 + \rho^{t+1}\theta_{2,j} + (1 - \rho^{t+1})\theta_{2,j} \\
&\geq Q_1 + \rho^{t+1}\theta_{1,j} + (1 - \rho^{t+1})\theta_{2,j} \\
&\geq Q_1 + \rho^{t+1}\theta_{1,j} + (1 - \rho^{t+1})\theta_{1,j} \\
&= Q_1 + \theta_{1,j} \\
&= U_j^t.
\end{aligned}$$

To show that consumption segregation increases over time, it is enough to show that the probability of two consumers taken randomly buying the same product decreases. This probability is given by $M_1^2 + M_2^2 = (1 - M_2)^2 + M_2^2$, which is clearly decreasing for $M_2 \in [0, 1/2]$.

The last point, meaning that the platform experiences rating inflation, is merely a Corollary of the fact that individual utilities increase, given my initial assumption of ratings as utilities.

$\blacksquare$

B "Observational" Learning: the Number of Reviews

Anecdotal evidence suggests that consumers have a preference for goods which are purchased by many other consumers. One possible explanation is that there are network effects of some sort. However, even absent these network effects there is a premium for high-sales products. Caminal and Vives [1996] provide an explanation: consumers use market shares as a signal of product quality. Powell et al. [2017] suggest that the quantity premium results from a psychological bias ("love of large numbers").

When it comes to consumer ratings, a similar empirical regularity is observed: consumers seem to have greater regard for products with a greater number of reviews (Floyd et al. [2014], You et al. [2015], Watson et al. [2018]). In this case, the puzzle is deeper since the explanation in Caminal and Vives [1996] seems to fail: if consumers have direct information regarding quality (average quality ratings) what additional information can the number of reviews contain?

My model – slightly adapted and expanded in here – provides one possible explanation based on two observations/assumptions: (a) consumers rate products according to their subjective experience, which in turn reflects both objective product quality and subjective fit between product characteristics and consumer preferences; (b) higher-quality products attract more consumers, including in particular consumer for whom the fit component is lower. A combination of (a) and (b) implies that rational consumers should use both average rating and number of ratings as a measure of quality.

To see how this may be the case, consider the following stylized model. There are n sellers, each of which sells one product. The product can be one of two types, type a and type b. In each of two periods, there is a measure 1 of consumers who are equally divided in terms of preferences for product type (that is, a measure $\frac{1}{2}$ has a preference for type a products. Let τ be the disutility from consuming a product of type different from the preferred type.

In addition to this element of horizontal product differentiation, we also assume that each product is characterized by vertical quality q, where quality units are the same as utility units.

Since the focus of the analysis is on learning about quality and match value, we assume that prices are exogenously given; and with no additional loss of generality, we assume prices are zero.[53]

[53] Alternatively, one may think of q as quality net of price.

There are two periods. First-period consumers do not have access to reviews. Second-period consumers, by contrast, have access to the reviews issued by first-period consumers.

Consider the case of first-period consumers. We assume they are randomly presented with one product and learn the product's quality q as well as its type t.[54] The decision problem is then easy: If the product's type is identical to the consumer's type, then the consumer makes a purchase if and only if $q > 0$. If, by contrast, the product and the consumer are of different type, then the consumer makes a purchase if and only if $q > \tau$. Finally, we assume that consumers leave an honest review after purchase, that is, the review score equals utility level q if there is a match of types and $q - \tau$ if there is no such match.

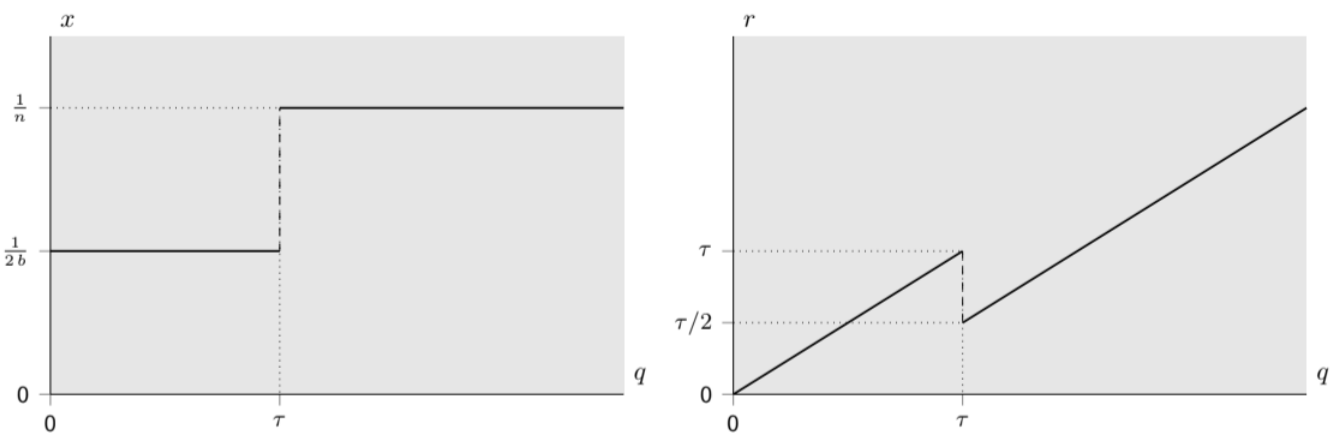


Figure 1: Relation between product quality q and first periods sales x (left panel) and between quality q and average consumer rating r (right panel)

The figure shows the relation between product quality q and first periods sales x (left panel) and between quality q and average consumer rating r (right panel). As the left panel shows, if quality q is greater than the τ threshold, then sales double, as the product attracts from both consumers with good fit and consumers with bad fit. As the right-hand panel shows, the relation between q and r is non-monotonic. This is because, as quality increases, the product attracts buyers for whom the fit component is lower, resulting in lower ratings.

Second period consumers do not observe the value of q (or don't need to make the investment into finding the value of q). Rather, they observe two summary statistics regarding first period purchase decisions: x and q. For simplicity, suppose that all buyers rate the product, so that x is both sales and number of reviews.

[54] The qualitative results remain valid if we consider an optimal search problem. This is a conjecture.

Similarly to first-period consumers, second-period consumers make a purchase if and only if net utility is positive. If there is a product fit, then the rule is $\mathbb{E}(q) > 0$. If there isn't a product fit, then the rule is $\mathbb{E}(q) > \tau$. Rational consumers "invert" the mapping on the right-hand panel of Figure , and so

$$\mathbb{E}(q) = \begin{cases} r & \text{if} \quad x \leq \frac{1}{2\,n} \\ r + \frac{1}{2}\tau & \text{if} \quad x > \frac{1}{2\,n} \end{cases}$$

The main point is that rational consumers use both r and x as measures of quality. This is not a bounded rationality issue or a network effects issue. It simply results from correcting for the subjective element in reviews and the "curse of success" in consumer ratings.

This simple model hints at an important interaction between "observational" learning and learning from reviews in markets with horizontal differentiation. I plan to expand on this line of research in the near future.

C Model Extensions

I now briefly discuss my model's robustness (or lack thereof) to changes in its core assumptions. Particularly, I have made three key Assumptions for my analysis: *i*) ratings are indicative of gross utility, that is, of the subjective quality of the product, and independent of price, as well as strategic considerations, *ii*) no self-selection at the writing stage, conditional on purchase: everyone that purchase a product rates it and *iii*) consumers are choosing between two options.

Inspired by both realism and the existing literature, I consider the following extensions.

C.1 Alternative Rating Behavior

C.1.1 Ratings as "Quality per Dollar"

In contemporary work, Luca and Reshef [2019] show empirically using *Yelp* data that a 1% price increase in price leads to a $3-5\%$ decrease in ratings. That is, with restaurants, consumers do not purely reward the quality of their meal, but rather its "quality per dollar". In my model, ratings are independent of price.

Empirically, it is important to recognize that price is much more salient in some markets than others. For instance, while reviews along the lines of "I had a good meal, but I found it too expensive" are common on *Yelp*, not a single one of the reviews in my dataset mentions this aspect. The argument is even more straightforward for movies, TV series and music: the former are uniformly priced, the latter usually come as part of a bundle with a fixed monthly submission fee, so that their marginal cost is effectively 0.

Moreover, on a theoretical level it is not clear why ratings should reflect prices, when these are public in the first place. Of course, the question becomes: "Are consumers aware that their predecessors' ratings included prices, and do they correct appropriately?" For simplicity I will abstract from this issue and assume they do.

Since I believe my results apply to the restaurant and hotel industries as well, it is worth mentioning robustness of my model to the alternative specification

$$\mathcal{R}_{ij} = \begin{cases} Q_i + \theta_{ij} - P_i & \text{if } i \in argmax\{\mathbb{E}(U_{1,j}), \mathbb{E}(U_{2,j})\}, \\ \varnothing & \text{otherwise.} \end{cases}$$

While the main predictions are unchanged, with niche and lower quality products being relatively advantaged in ratings, some of the managerial implications change dramatically. In my baseline model, prices are a *matching device*: a higher price selects a more devoted set of fans. This prediction is robust; as is the fact that naïve consumers are going to overestimate the (now net) quality of the product, compared to their rational peers. Both follow from the fact that self-selection was predicated on my price in my initial specification as well. However, price now has a direct negative effect, too, altering the tradeoffs.

Therefore, predictions on the sign of $\frac{\partial \mathbb{E}(\mathcal{R}_i^{t-1})}{\partial P_i}$ become more ambiguous, and depend on the specific functional forms. When high prices are more punished in ratings than they help selecting the ideal consumers, my managerial advice becomes much more similar to that of Luca and Reshef [2019]: firms should be vary of increasing their prices to cash in on their reputation, as this might hurt their future one and prove detrimental to their long-term success.

C.1.2 Strategic Ratings

Another interesting deviation from my model is that of strategic rating behavior. That is, when leaving ratings, consumers' main motivation might be to influence their successors' choices, rather than posting a subjectively accurate rating. Generally, consumers motivated by persuading their peers will not rate truthfully.

For example, consider a consumer who believes that a product is of good quality (say, 4 out of 5), and before posting, notices that the product currently has an average rating of 3.5. Then, her best response is to inflate her rating to 5, to get the *ex-post* average rating closer to her belief, 4. This behavior has been empirically documented by Wu and Huberman [2008]).

That is, for a product of quality Q_i for which she has taste θ_{i,j^*}, she might be trying to maximise the strategic (S) loss function[55]

$$L^S(Q_i, \theta_{i,j^*}) := -\big(Q_i + \theta_{i,j^*} - \mathbb{E}(\mathcal{R}_i | \mathcal{R}_{i,j^*})\big)^2$$

instead of the purely individual (I) one

$$L^I(Q_i, \theta_{i,j^*}) := -\big(Q_i + \theta_{i,j^*} - \mathcal{R}_{i,j^*}\big)^2.$$

The latter clearly leads to truth-telling, $\mathcal{R}^I_{i,j^*} = Q_i + \theta^*_{i,j}$, and can be seen as a formal microfoundation for my model. An in-depth study of social learning with strategic rating

[55] The results continue to hold when she maximises a weighted average of individual and collective precision.

behavior is beyond the scope of this paper, and seems a promising area for future research (as also suggested by Acemoglu et al. [2017]). Here, I will only add two observations.

The first one is that $L^I(Q_i, \theta_{i,j}) \approx L^S(Q_i, \theta_{i,j})$ whenever the number of ratings the product had already received, $\mathcal{N}(\mathcal{R}_i)$, is large, and the consumer does not receive a disproportionate weight. When reviewing a book with two million reviews, there is very little scope for strategic behavior, as every consumer is essentially atomistic. This is the case on *Goodreads*, where every book has several thousands (if not millions) of reviews. Thus, this concern is, at most, limited to very few, obscure, books, and even for them, to their initial ratings.

Second, notice that for each $j^* \in \mathcal{J}_i$, t is straightforward to sign the difference between individual and strategic reviews, $\mathcal{R}^I_{i,j^*} - \mathcal{R}^S_{i,j^*}$:

$$\mathcal{R}^I_{i,j^*} < (>) \mathcal{R}^S_{i,j^*} \;\Leftrightarrow\; \mathbb{E}(\mathcal{R}_i) < (>) Q_i + \theta_{i,j^*} \;\Leftrightarrow\; \mathbb{E}(\theta_{ij} | j \in \mathcal{J}_i) < (>) \theta_{i,j^*}.$$

This means that, even if consumers were strategic, this would benefit in particular products of lower quality, and niche designs, that only appeal to few consumers. To see this, observe that, with strategic consumers, products with very positive (negative) ratings will mostly incentivize unsatisfied (satisfied) consumers to deviate from truthful ratings, compressing ratings and underestimating quality differences.

C.1.3 Social Influence

Social influence is "the change in behavior that one person causes in another, intentionally or unintentionally, as a result of the way the changed person perceives themselves in relationship to the influencer". In the context of experienced utility, and particularly of consumer reviews, this translates into "biasing the judgement of an experience – and, thus, adapting one's rating – in the direction of what previous consumers have reported".

For instance, if every consumer in the previous generation has reported that a product is fantastic, future consumers might perceive it as such even when they would not have if they were to consume it in isolation. That is,

$$\frac{\partial U^t_{ij}\big(Q_i, \theta_{i,j}, \mathbb{E}^{t-1}(\mathcal{R}_i)\big)}{\partial \mathbb{E}^{t-1}(\mathcal{R}_i)} > 0.$$

For example, using a large scale field experiment, Muchnik et al. [2013] demonstrate that randomly manipulating the first upvote or downvote received by a user post on an online forum influences its long-term upvotes to downvotes ratio, while Jacobsen [2015] shows that when famous beer bloggers review a beer more positively or negatively than the average of consumers, future consumer ratings shift in the direction of the critics' opinion. Notice how this is opposite to what described in C.1.2, in which consumers effectively look as contrarians, despite lacking of social image concerns.

My model is *not* robust to social influence, and in fact generates opposite predictions to it. Clearly, the presence of social influence leads to *winners take all dynamics*: better ratings today translate into (more and) better ones tomorrow. Particularly, the opinions of particularly influential members (such as *New York Times* book critic Michiko Kakutani, or the juries of important literary prizes) should be able to sway not only readers' choices, but also their very perception conditional on that.

I believe a variety of my empirical findings, as well as those of Kovács and Sharkey [2014], should offer substantial evidence that social influence is not prevalent in this context and, if anything, excessively high previous ratings and exposure often end up *hurting* a product's future reputation. In light of these empirical findings, I see the fact that my model offers opposite predictions to those of social influence theory as one of its key features, rather than a limitation.

C.2 Self-Selection Into Leaving Reviews

In my model, everyone leaves a review upon purchasing a product. In reality, very few consumers leave reviews: a variety of surveys estimate this percentage to lie between 1% and 5%, depending on the market.

Given the severity of this additional source of self-selection, it is important to understand its nature, and thus its implications for my main findings. In line with the empirical literature, I will focus on a specific form of self-selection: *extremity bias*.[56]

It is well studied, observationally and experimentally (Brandes et al. [2018]), that reviews often display a bimodal distribution, suggesting that consumers with strong (either positive or negative) feelings towards the product are more likely to post a rating.

[56] A different type of self-selection at the writing stage would be one based purely on the product's characteristics (e.g., its popularity) more so than on the consumer's satisfaction with it. In my data, by comparing the number of numerical ratings (which take a few seconds to post) to written reviews (which take substantially longer), I can get a sense of which products are more likely to motivate consumers to spend time reviewing them. I find that the ratio of reviews to ratings is lower for products that have more reviews. This suggests that consumers might not find it worthy to invest their time in reviewing a product that has already received a lot of attention from their peers, in line with a model of strategic ratings motivated by the desire to influence others.

I believe this bias is less severe on *Goodreads*, given its users have joined the platform at least in part to get personalised book recommendations, and these improve the more information the users shares about what they are reading and enjoying. In fact, we see a high fraction of non-extreme ratings, and for most products the combined share 5's and 1's lies below 50%, which is close to the random benchmark of 40%. Moreover, there are usually many more 5's than 1's, and this might be due to consumers finding products they legitimately love more so than to extremity bias.

Even if this bias were prevalent, my model is fairly robust to it. Note that I model utilities, and thus ratings, as linear in both quality and match value. Because of this, simply assuming that consumers in both tails (say, consumers that are either below the $10th$ percentile or above the $90th$ in their idiosyncratic taste for the product) leave reviews does not change any of the conclusions. That is, whenever the projection of $\mathcal{J}_i$ onto the support of θ_i is an interval for each value of θ_{-i} – as is the case under mild informational assumptions – we have

$$\mathbb{E}(\theta_{ij}|\theta_{ij} \in \mathcal{J}_i) = \frac{1}{2}\mathbb{E}(\theta_{ij}|\theta_{ij} \in \mathcal{J}_i^{10-}) + \frac{1}{2}\mathbb{E}(\theta_{ij}|\theta_{ij} \in \mathcal{J}_i^{90+}).$$

A more complex case is the one in which love and hate for the product are measured in absolute terms, that is, consumer j leaves a review for product i when either $U_{i,j} > \bar{U}$ or $U_{ij} < \underline{U}$, for two consumer and product independent thresholds $\underline{U} < \bar{U}$. Under these assumptions, the average ratings of low quality products would be downward biased, while the opposite is true for products of high quality. Similarly, niche products would receive more positive ratings than their mainstream counterparts, provided that the right crowd finds them. When niche products are also of lower quality (Johnson and Myatt [2006], Bar-Isaac et al. [2012]), the conclusions are ambiguous.

C.3 Large Product Space

For tractability as well as ease of exposition, my model deals with only two products. But of course, most markets – and particularly those in which consumers rely on online reviews the most – feature a vast number of products.

The first effect of a large product space, independently on word of mouth, is an increase in the share of niche products. Intuitively, when competing against many alternatives, it is better to be someone's first choice than everyone's second or third. Thus, a strong taste for the product is needed, leading to more polarizing designs. This is true for every product apart from those of extremely high quality (see Johnson and Myatt [2006], Bar-Isaac et al.

[2012]).

This means that: *i*) the reviews of niche products will be inflated as usual, but (aside from quality differences) relatively fair, but *ii*) the bias between the ratings of the very high quality, mainstream products and the much lower quality, niche alternatives is larger than it is the case in the two products model. Formally, denoting by i^* a niche product (and by $-i^*$ its only alternative in the two products case), we have

$$\mathbb{E}\big(\theta_{i^*,j}|Q_{i^*}+\theta_{i^*,j}-P_{i^*}=\max_{i\in\mathcal{I}}\ Q_i+\theta_{ij}-P_i\big)>\mathbb{E}\big(\theta_{i^*,j}|Q_{i^*}+\theta_{i^*,j}-P_{i^*}>Q_{-i^*}+\theta_{-i^*,j}-P_{-i^*}\big),$$

namely, the conditional idiosyncratic taste is stronger in the multiple products case. Moreover, what is key to my claim is that the gap between the left and the right hand side is higher for niche products than it is for the mainstream ones: again, this follows from the greater variance of polarizing products' shocks, leading to greater upside for their truncated distributions.

D Data and Descriptive Statistics

D.1 Goodreads

D.1.1 Goodreads Data Examples

Figure 2: Typical Book Page

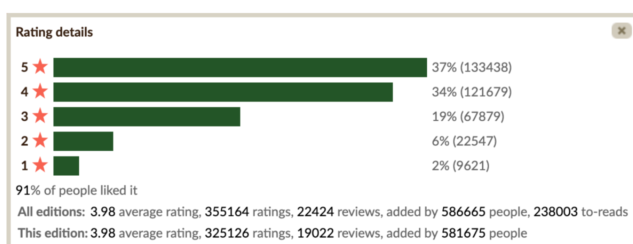


Figure 3: Advanced Book Statistics

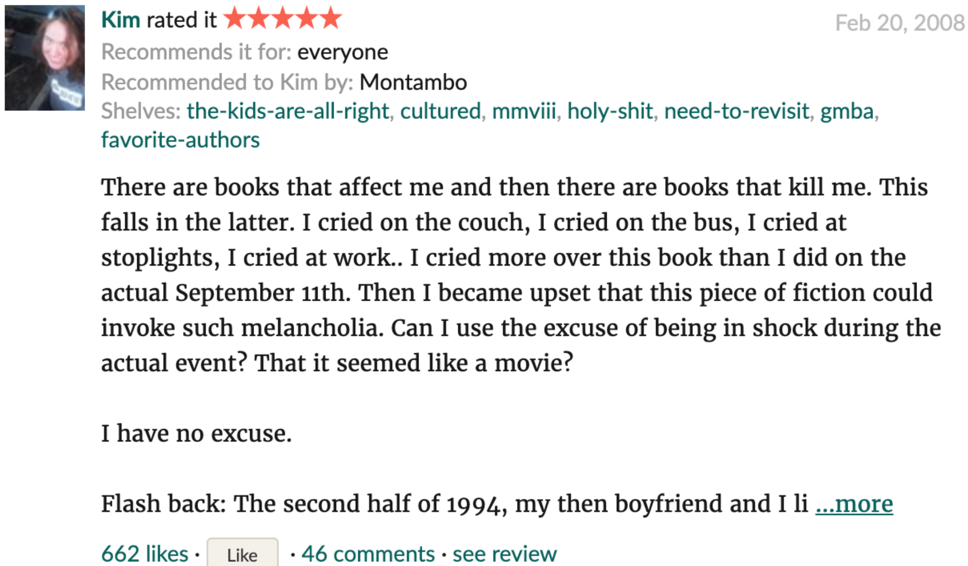
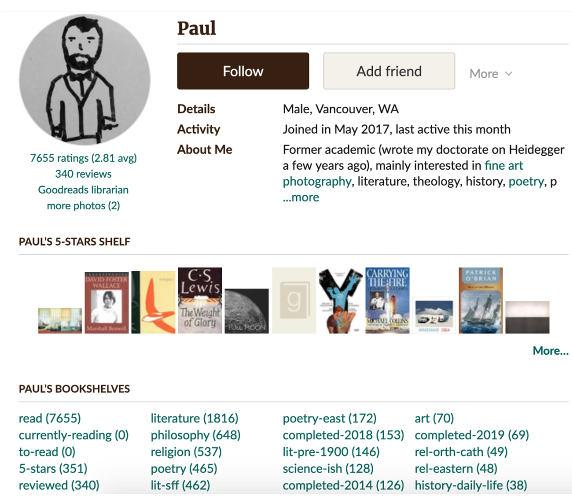
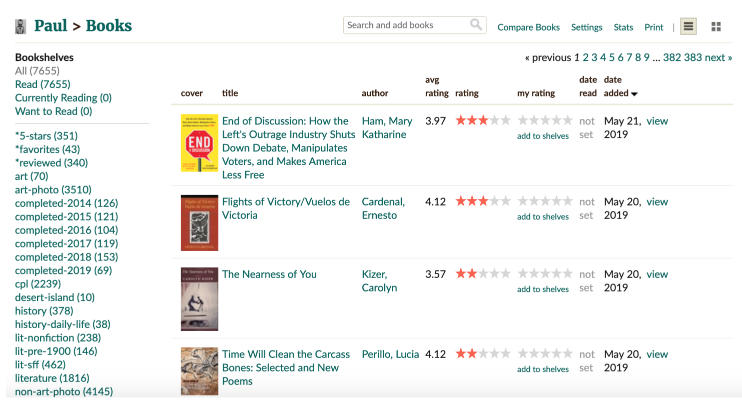


Figure 4: Typical Book Review



(a) Demographics and Reading Information



(b) Ratings Data

Figure 5: A Typical *Goodreads* User

D.1.2 Goodreads Descriptive Statistics

Table 1: Descriptive Statistics for All Scraped *Goodreads* Books

Statistic	N	Mean	StDev	Min	25th Pctl	75th Pctl	Max
Number of Editions	3,128	62	244.581	1	4	48	5,578
Number of Ratings	3,128	82,221	321,458	0	282.5	32,236.2	5,887,954
Number of Reviews	3,128	3,504	10,380	0	68	2,137	162,401
To Read	3,128	33,806	90,249	0	1,632.8	22,067.5	1,143,131
Average Rating	3,128	3.987	0.311	2	3.8	4.19	5
Number of Fives	3,128	34,516	158,233	0	93.8	11,451	3,780,823
Number of Fours	3,128	26,437	96,283	0	101	11,014	1,718,918
Number of Threes	3,128	14,633	51,882	0	50.8	6,488	893,665
Number of Twos	3,128	4,403	18,068	0	13	1,642	487,003
Number of Ones	3,128	2,229	12,762	0	4	592	500,546

Figure 6: The Distribution of Average Ratings

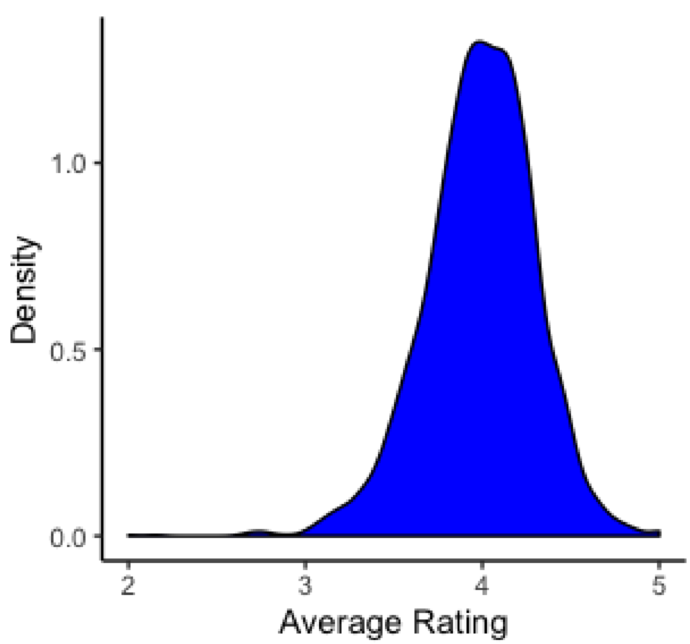


Figure 7: The Distribution of the Variance of Ratings

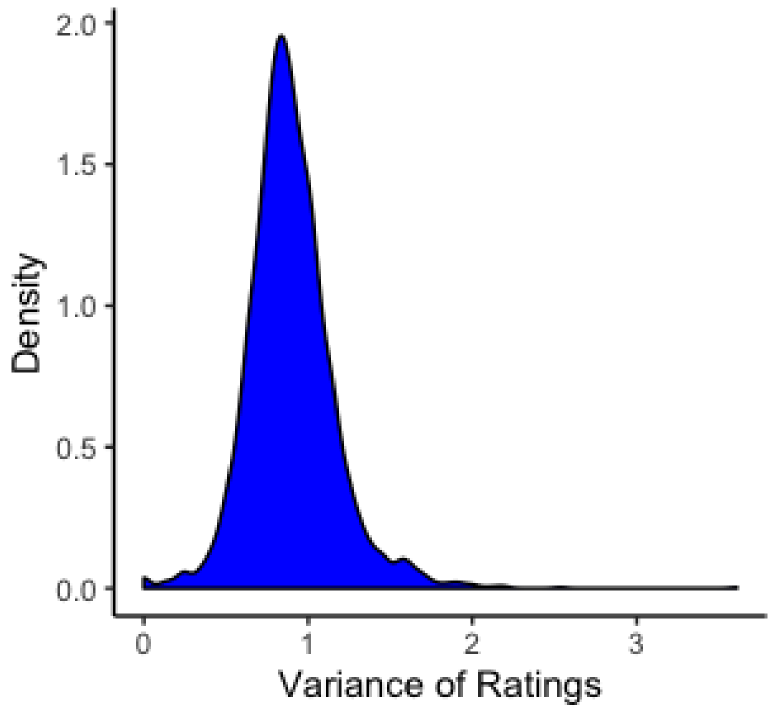


Table 2: Goodreads Users: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Users Statistics					
Number of Reviews	100,100	47.843	185.402	0	3,158
Number of Ratings	100,100	300.717	177.923	0	7,592
Average Rating	100,100	3.936	0.725	3.14	5
Active	100,100	0.848	0.359	0	1
User Genre Count	100,100	24.716	21.510	1	59
Leniency	100,100	0.001	0.641	−3.61	2.45
Years Active	100,100	4.511	2.961	0	12
Books Statistics					
Average Rating	100,100	4.010	0.411	1	5
Number of Ratings	100,100	249,782	694,061	0	5,923,980

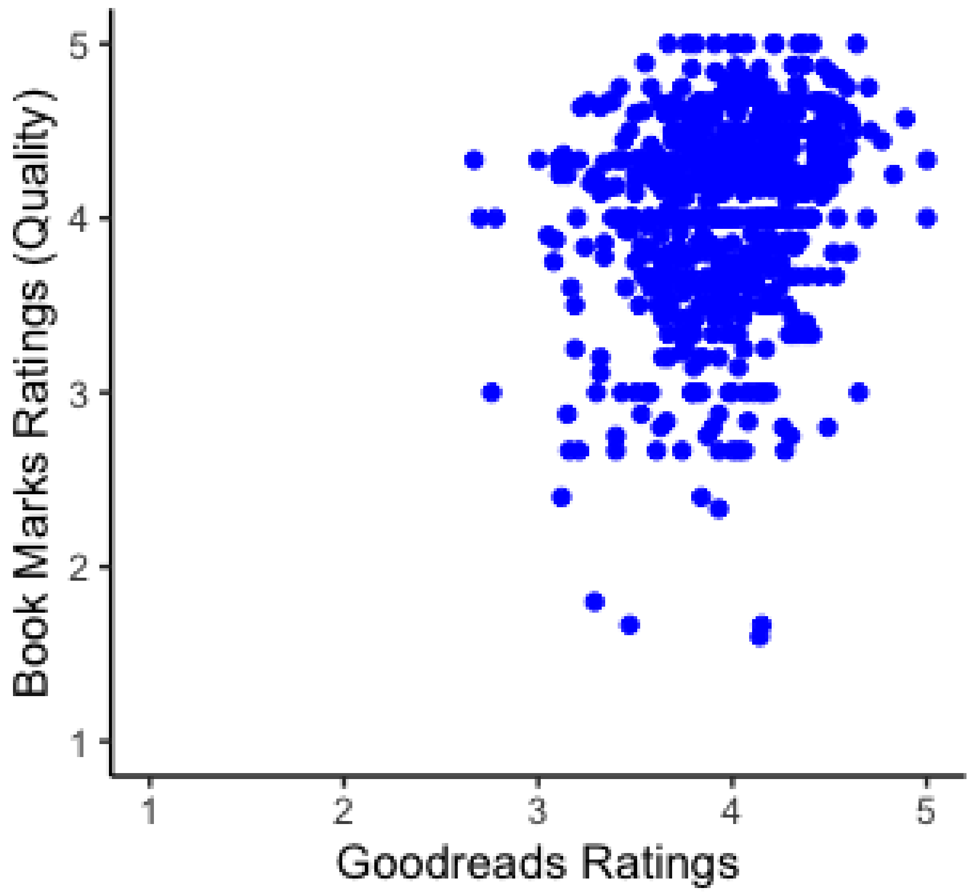
D.2 Book Marks

Figure 8: Typical book page on Book Marks

E Results

E.1 Goodreads Relatively Penalizes High Quality Products

Figure 9: A Comparison of *Goodreads* and *Book Marks* Ratings



E.2 The Variance of Ratings

Figure 10: Number vs Variance of Ratings

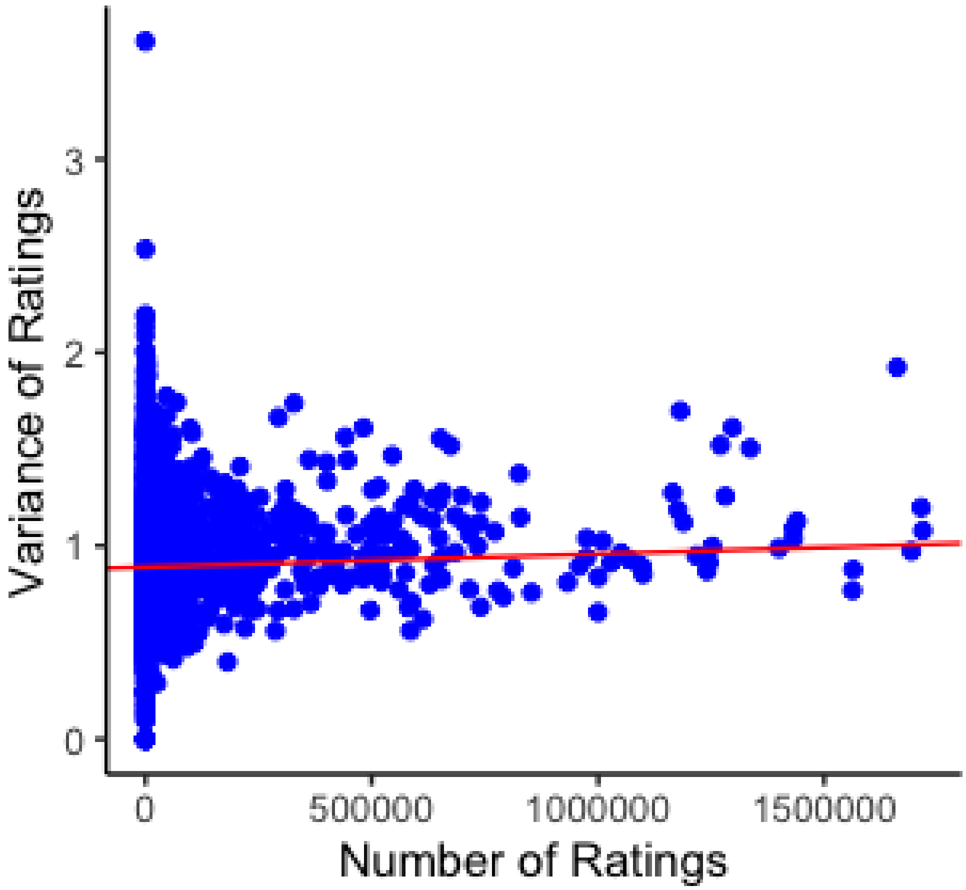


Table 3: Determinants of Variance in Ratings

	Dependent variable:
	Variance of Ratings
Average Rating	-0.498^{***}
	(0.012)
Number of Ratings	0.0001^{***}
	(0.000)
Unusual Genre	0.008
	(0.009)
Year	-0.0001^{***}
	(0.00001)
Major Award	0.020
	(0.012)
Number of Editions	0.00005^{***}
	(0.00002)
Observations	3,128
R^2	0.363
Adjusted R^2	0.361
Residual Std. Error	0.211 (df = 3121)
F Statistic	295.870^{***} (df = 6; 3121)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 11: Variance of Ratings: LASSO Coefficients

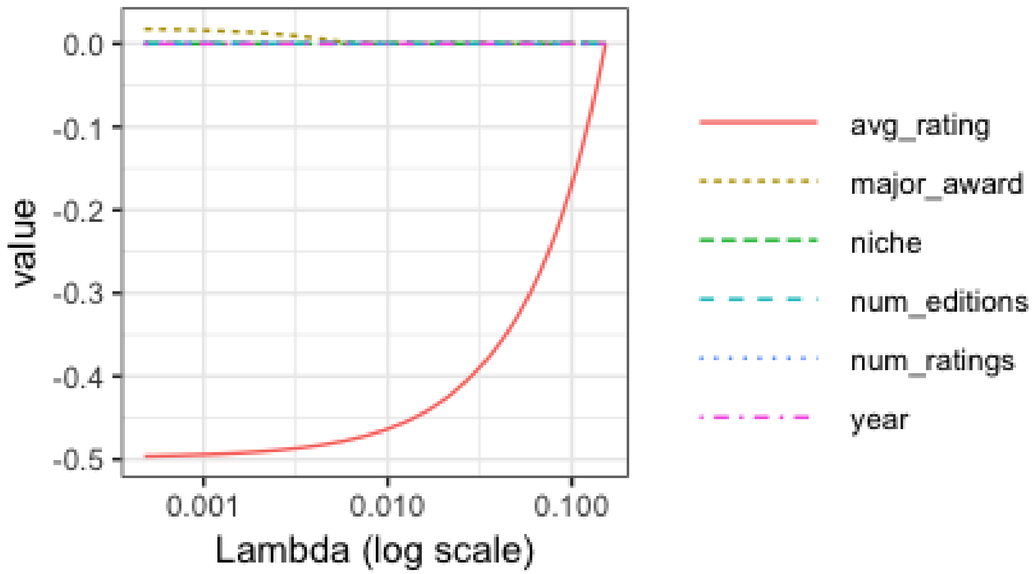
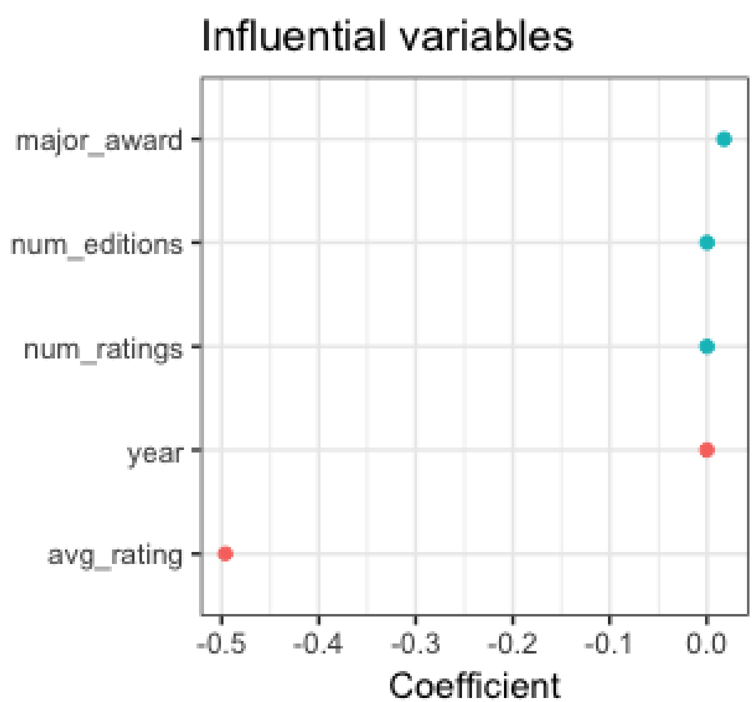


Figure 12: Variance of Ratings: Influential Variables



E.3 Goodreads Users Behavior

E.3.1 Early vs Late Behavior

Figure 13: Empirical CDF of Book Popularity, Early vs Late

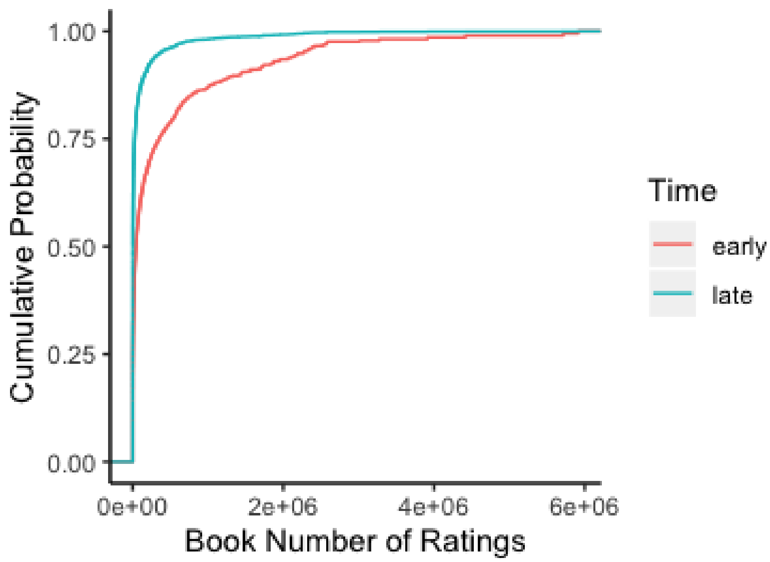


Figure 14: Empirical CDF of Book Popularity, Early vs Late

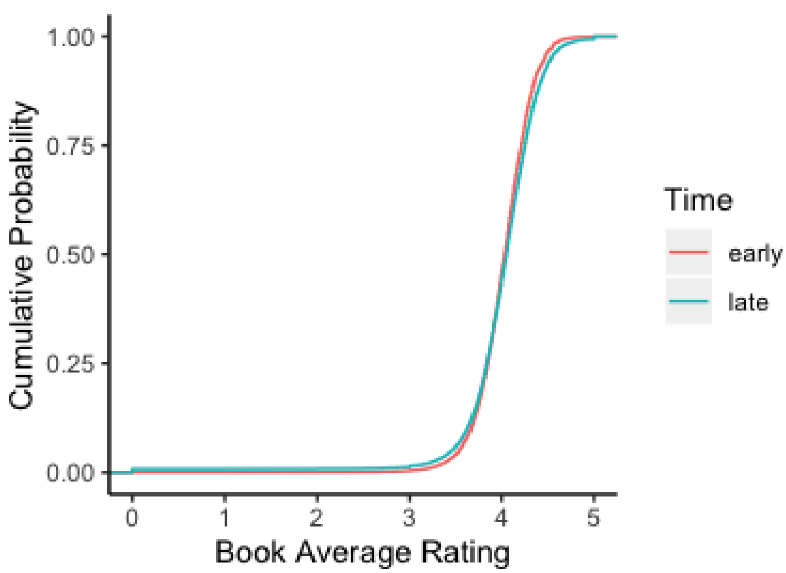


Table 4: Leniency for Different Books

	Dependent variable:	
	Leniency	
	OLS	*User Fixed Effects*
	(1)	(2)
Late	-0.082^{***}	-0.076^{***}
	(0.004)	(0.004)
Unusual Genre	-0.092^{***}	-0.084^{***}
	(0.017)	(0.016)
Book Number of Ratings	-0.001^{***}	0.000
	(0.000)	(0.000)
Late*Unusual Genre	0.102^{***}	0.074^{***}
	(0.022)	(0.021)
Constant	0.046^{***}	
	(0.003)	
Observations	100,100	100,100
R^2	0.004	0.135
Adjusted R^2	0.004	0.126
Note:	*p<0.1; **p<0.05; ***p<0.01	

E.3.2 Cross Sectional Analysis

Figure 15: User Ratings vs Book Popularity

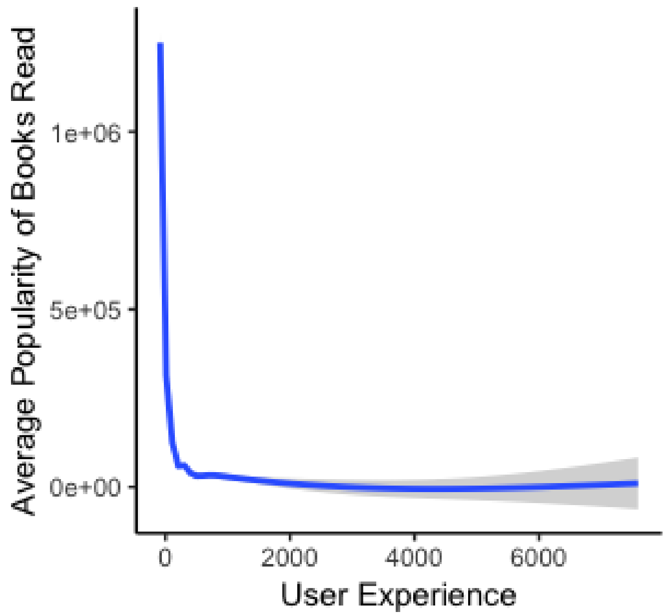


Figure 16: Years Active vs Book Popularity

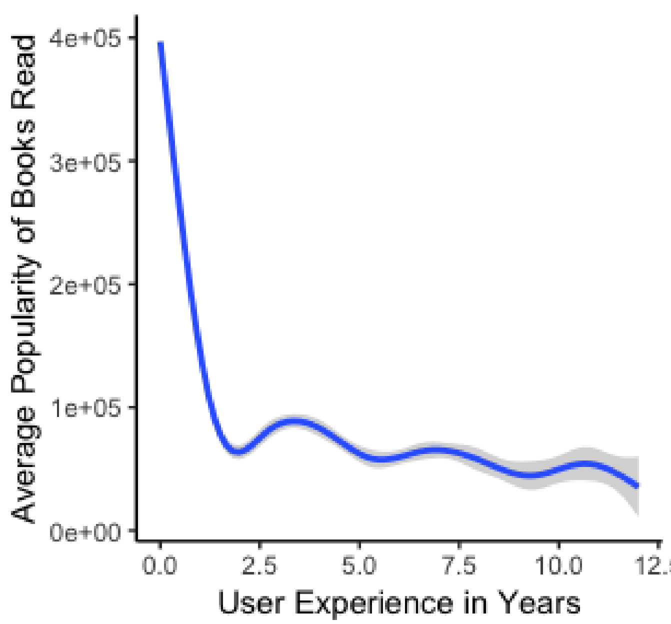


Figure 17: User Ratings vs Quality

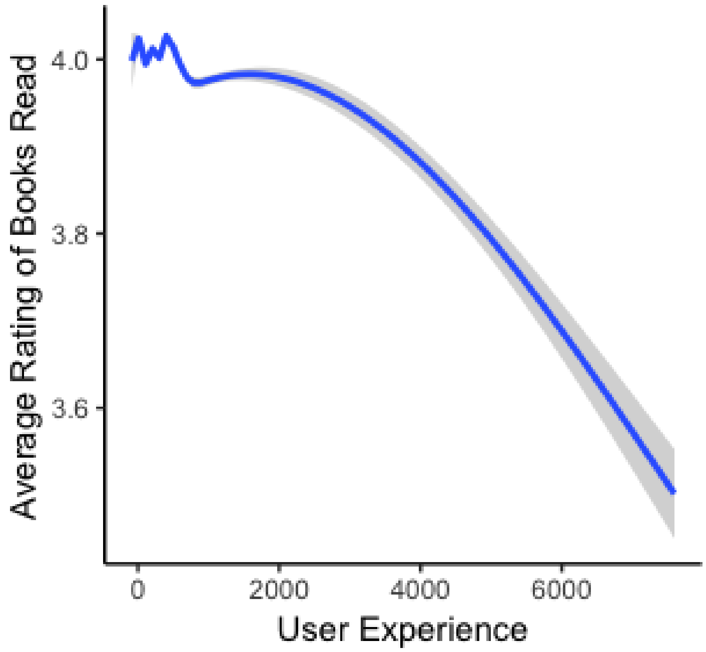


Figure 18: Years Active vs Quality

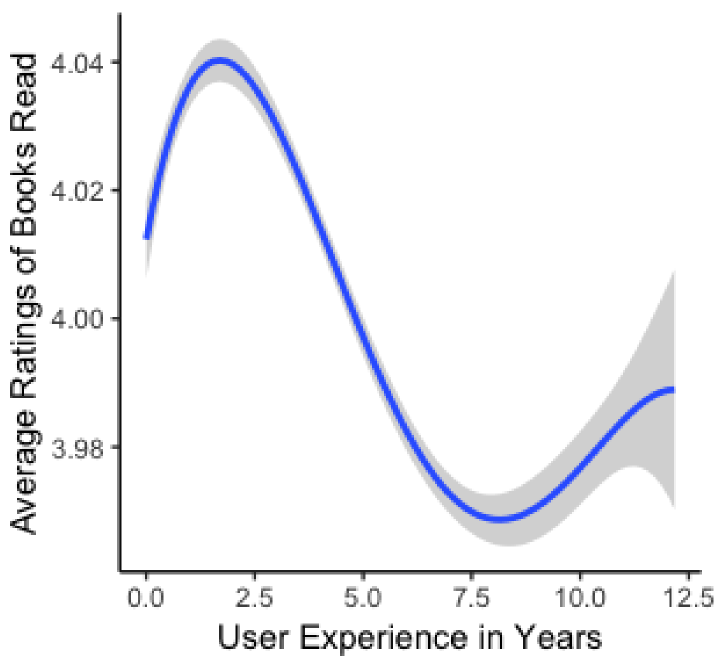


Figure 19: User Ratings vs Share of Best Selling Books

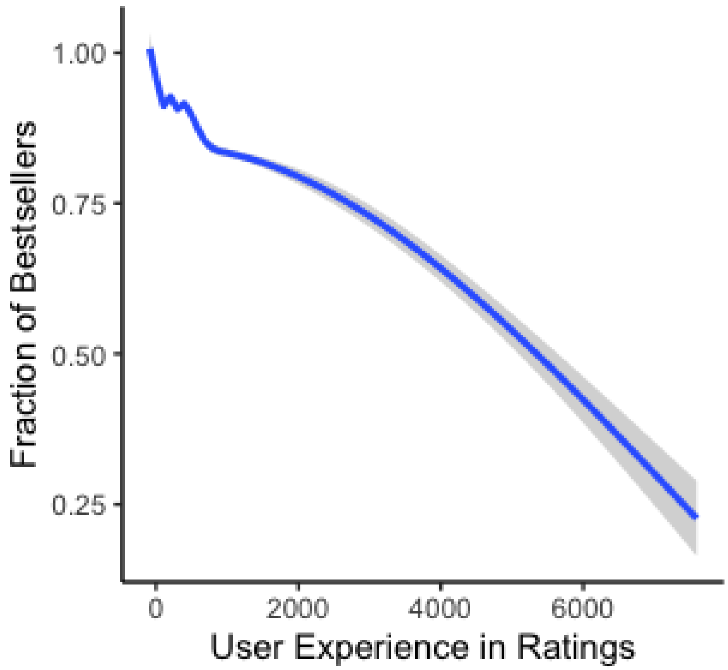


Figure 20: User Ratings vs Share of Obscure Books

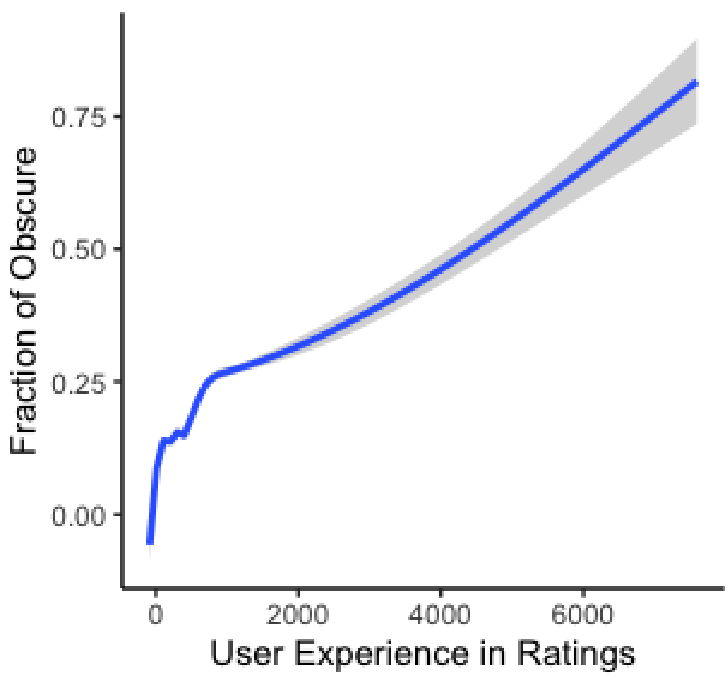


Figure 21: User Ratings vs Unusual Genre Books

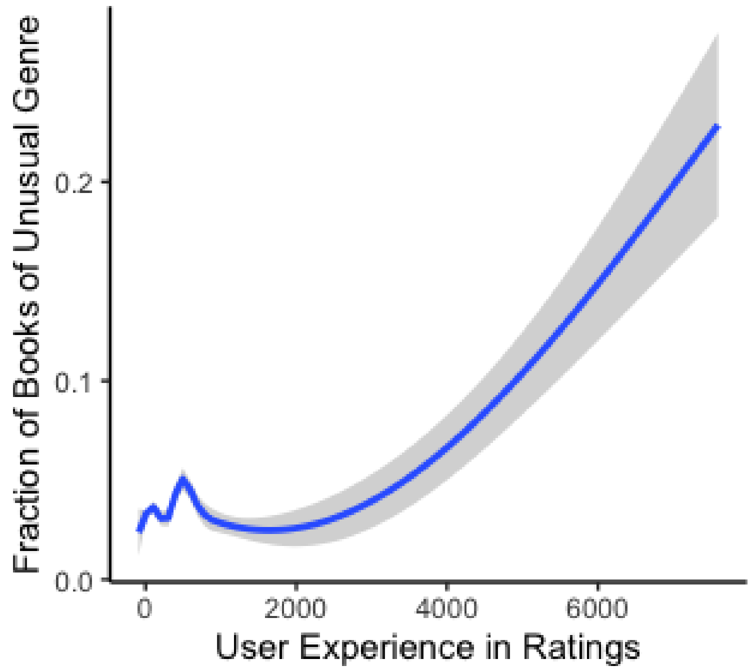


Figure 22: User Ratings vs Leniency

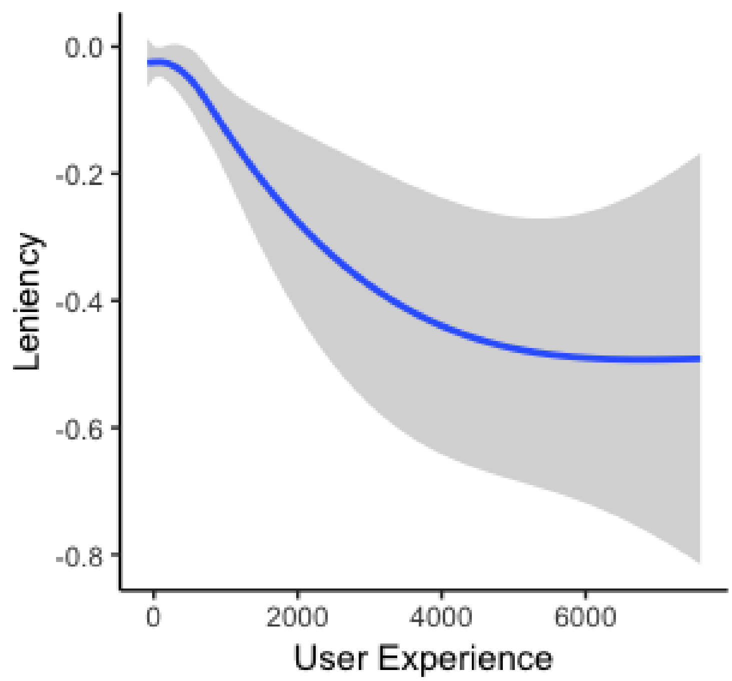


Table 5: Leniency with Obscure Books

	Dependent variable:	
	Leniency	
	OLS	*User Fixed Effects*
	(1)	(2)
Experience in Ratings	-0.0002^{***}	-0.0001^{***}
	(0.00001)	(0.00001)
Obscure	-0.025^{***}	-0.061^{***}
	(0.007)	(0.007)
Experience in Rating * Obscure	0.0001^{***}	0.0001^{***}
	(0.00001)	(0.00001)
Constant	0.024^{***}	
	(0.002)	
Observations	100,100	100,100
R^2	0.006	0.134
Adjusted R^2	0.006	0.125
Residual Std. Error	0.639 (df = 100096)	0.599 (df = 99097)
F Statistic	189.983^{***} (df = 3; 100096)	
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

E.4 Demand for Information

Table 6: Reviews Numerical Scores

	Dependent variable:
	Number of Stars
Average Rating	1.146^{***}
	(0.010)
Default	-0.033^{***}
	(0.008)
Oldest	0.138^{***}
	(0.006)
Constant	-0.683^{***}
	(0.039)
Observations	140,215
R^2	0.094
Adjusted R^2	0.094
Residual Std. Error	1.026 (df = 140211)
F Statistic	$4{,}823.026^{***}$ (df = 3; 140211)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Reviews Usefulness

	Dependent variable:
	Number of Likes
Number of Reviews	0.001^{***}
	(0.00001)
Review Has Recommend	4.139^{***}
	(0.623)
Character Length	0.011^{***}
	(0.0001)
Number of Stars	-0.679^{***}
	(0.132)
Constant	-1.821^{***}
	(0.550)
Observations	140,215
R^2	0.096
Adjusted R^2	0.096
Residual Std. Error	53.084 (df = 140210)
F Statistic	$3{,}717.256^{***}$ (df = 4; 140210)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 8: Reviews Usefulness - Unknown Books

	Dependent variable:
	Number of Likes
Average Rating	-0.196
	(0.221)
Number of Reviews	0.013^{***}
	(0.004)
Recommended for Some	4.731^{***}
	(1.014)
Character Length	0.001^{***}
	(0.0001)
Number of Stars	0.312^{***}
	(0.068)
Constant	-0.291
	(0.856)
Observations	14,005
R^2	0.053
Adjusted R^2	0.052
Residual Std. Error	7.624 (df = 13999)
F Statistic	155.733^{***} (df = 5; 13999)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Reviews Usefulness - bestsellers

	Dependent variable:
	Number of Likes
Average Rating	12.057**
	(5.141)
Number of Reviews	0.001***
	(0.0001)
Recommended for Some	8.652
	(7.086)
Character Length	0.037***
	(0.001)
Number of Stars	−6.101***
	(1.095)
Constant	−48.304**
	(19.605)
Observations	14,010
R^2	0.147
Adjusted R^2	0.147
Residual Std. Error	144.696 (df = 14004)
F Statistic	483.994*** (df = 5; 14004)
Note:	*p<0.1; **p<0.05; ***p<0.01