WRITING SAMPLE: ECONOMICS HONOURS DISSERTATION

SELF-SELECTION AND SOCIAL INFLUENCE BIASES IN REVIEWING BEHAVIOUR AN EMPIRICAL STUDY OF REVIEWS ON AMAZON.COM

Supervisor: Dr Jovan Vojnovic

The University of Edinburgh
School of Economics
April 2021
Word Count: 9966

Despite a challenging year, during a global pandemic, I was grateful to have by my side my supervisor Dr Jovan Vojnovic, who provided much needed academic and emotional support. Further gratitude goes to Dr Timothy Cannings, Artem Yacenko and Filip Pilarcik for discussing ideas with me. Lastly, appreciation goes to my brother, Andreas Aristidou, for his feedback and Dimitris Christodoulou for his assistance when coding in programming languages.

Abstract

This dissertation is concerned with the joint acknowledgement that online feedback communities are crucial in building trust and facilitating choice in online commerce, while behavioural biases affecting reviewing behaviour distort the accuracy of gathered feedback. This significantly diminishes the usefulness of feedback systems for their intended purpose and may provide misleading inferences on product quality. Through a series of empirical analysis, including sentiment analysis, on a sample of reviews from Amazon.com, we examine how self-selection and social influence biases give rise to asymmetrically bimodal sampling distribution of ratings, closely resembling a J-curve shape. The results reveal that infrequent reviewers are driven by reciprocity, indicating that a greater proportion of their ratings are extreme with either 1* or 5* score. We further observe that reviewing behaviour is likely influenced by existing reviews through the mechanisms of conformity and anchoring. Upon observing the potential drawbacks of feedback communities, we recommend a series of improvements that can correct biased behaviour in reviewing. We conclude the study by discussing the limitations of our analysis and recognising the potential for further research.

Contents

1 Introduction

Reviews posted in feedback communities play an important role in shaping customers' purchasing decisions in online marketplaces. They help buyers decide which product to choose and provide a system for building trust between buyers and sellers. To illustrate why this is especially important in online marketplaces, consider the physical marketplaces first. In physical marketplaces, reputation and quality are discovered either through repeated interactions or by word-of-mouth. Recommendations from family, tastings of various coffee sorts in the favourite cafe or a friendship with the barista all provide valuable information that shapes consumer choices. Conversely, online marketplaces suffer from anonymity and distrust between sellers and buyers, which constitutes a fundamental barrier to online markets. Online feedback systems attempt to overcome this problem by leveraging massive communities of past buyers, creating trust between sellers and buyers in an otherwise strangers' marketplace of often one-time interactions. Online word-of-mouth is regarded as a powerful marketing and advertising resource affecting sales and new product adoption, as Babić Rosario *et al.* (2016) and López and Sicilia (2013) show. Reviews, thus, play a vital role in ensuring the smooth functioning of online markets, which is demonstrated by the success of platforms such as Amazon and eBay, which place reviews at the forefront of the buyer's decision-making process.

Besides the advantages of feedback communities, some inherent flaws could ultimately offer misleading information. Firstly, engagement in feedback communities is entirely voluntary. Rational users have no clear incentive to provide feedback on their experiences. Those who do so are mainly driven by other-regarding preferences such as altruism, reciprocity or a need for attention. Secondly, feedback is self-reported, which implies that the review content is entirely subjective, heterogeneous, and may not represent the user's experience. Consider Alex, a prospective buyer in a physical marketplace within his community. To gather information on a specific seller, he could ask Bob, a friend who had a previous interaction with that seller. Bob, characterised as a homo-economicus, has an incentive to provide accurate feedback because he cares for Alex's utility and because he might need similar advice from Alex in the future. In online marketplaces, many users are simply free-riders, Kim and Walker (1984). They use the information to make better-informed choices but decide against providing information, as doing so involves a cost of time and effort with no apparent benefit. In physical marketplaces, both flaws mentioned above have a reduced effect thanks to community members such as friends. Online, the information available is overwhelming, and

the decision of whom to trust is fundamentally non-trivial. Therefore, feedback communities must decide how to summarise the aggregation of information they get from reviewers to facilitate the transfer of information. Kohei Kawamura suggests that we should not take information, such as reviews, at face-value. Instead, feedback communities need to find an optimal way to decode information, Kawamura (2011) and Kawamura (2013).

The most common way to summarise information from reviews is to calculate the average rating, using the individual rating corresponding to each review. However, conventional statistics argues that to make an inference on the mean, the underlying sampling distribution of the data must be unimodal or symmetrically bimodal. According to the central limit theorem, as explained in Anderson (2010), we expect that, under classical assumptions, any parental distribution will approximately converge to a normal distribution. In feedback communities, given the vast amounts of data available, this condition is expected to be satisfied. However, in section 6.1 we prove, using statistical tests, that the sampling distribution of ratings instead follows a non-symmetric bimodal distribution closely resembling a J-curve. Specifically, under the J-curve distribution, the middle values of the distribution (2*, 3*, 4*) are less dense than the tails of the distribution (1*, 5*), and the high-end tail (5*) is highly denser than the low-end tail (1*). The purpose of this paper is to examine whether and to what extent self-selection and social influence via quality disconfirmation can explain the observed sampling distribution of ratings.

1.1 Hypotheses

Self-selection bias can be divided into two types, *purchasing self-selection* and *engagement self-selection*. Purchasing self-selection bias portrays that customers purchase products, which they expect to like. Specifically, the prior belief that they will enjoy the product is not uniformly distributed; rather, it is amplified on the higher end of the rating scale. For example, vegan customers are more likely to purchase and thus enjoy tofu. Purchasing self-selection bias can explain why ratings are inflated and why the median is 5* instead of 3*. We will not focus on examining this particular bias in this dissertation; instead, we will closely analyse the effect of engagement self-selection bias. This bias refers to the observation that not all types of customers will engage in feedback communities equally often. While many customers never review, another group of customers contribute to feedback communities regularly, motivated by altruism and social awareness. However, we observe that most reviewers do not engage in feedback communities

regularly and, thus, we cannot allocate them into any of the previously specified groups. This violates the statistical axiom of sampling randomness. In this paper, we refer to the reviewers who often review as frequent and those who occasionally review as infrequent.

Hypothesis I: Reciprocity, rewarding sellers with a 5 review or punishing them with a 1* review, is a major driver of the J-curve ratings sampling distribution. Such behaviour is evident because infrequent reviewers usually only review their highest emotion-driven experiences, rated with an extreme rating.*

Social influence bias illustrates that reviewers who are about to post a review are affected by the already published reviews. People are generally uncomfortable deviating from the norm and are therefore more likely to review similarly to previous customers. This bias violates the statistical axiom of independence. This paper examines how reviewers are affected by *quality disconfirmation*, which captures the difference between expected and actual product quality.

Hypothesis II: Social influence affects reviewers via the magnitude of quality disconfirmation, conformity and anchoring biases.

The paper is organised as follows. We initiate our analysis by discussing relevant academic literature and theoretical underpinnings in section 2 and section 3. In section 4, we highlight the characteristics of the data, describing key variables and summary statistics. Section 5 introduces models and techniques used for the empirical analysis and are applied when discussing the results and their implications in section 6. We summarise our findings in section 7, drawing attention to the limitations of our analysis and the potential for further research.

2 Literature Review

We discuss relevant academic papers, examining why users choose to review and what inference we can gather from feedback communities. Further, we explore biases in reviewing behaviour and potential methods for removing such biases.

2.1 Why people rate?

Lafky (2014) uses theoretical models to identify the motivations driving consumers to review and examine how these factors create distortions in reviewing behaviour. The paper finds evidence that the cost of providing a review determines reviewers' behaviour, which is demonstrated by the fact that low-quality and high-quality sellers tend to be reviewed more than moderate-quality sellers. This finding is consistent with our results, where the rating distribution tails (1* and 5*) are denser than the middle of the distribution. Lafky (2014) only examines the distinction between reviewing and not reviewing, neglecting the possibility of a non-perfect mapping between experience and rating. Their model, thus, allows users to decide not to review, but conditional on reviewing, users rate without a bias. This dissertation instead examines biases in the mapping between experience and reviewing. Shen *et al.* (2013) study how reviewers compete for attention in online marketplaces. The paper follows the academic literature on understanding how social motivations affect reviewers' behaviour. They find that reviewers with high reputation tend to offer more modest ratings and use fewer negative words than infrequent reviewers. According to Shen *et al.* (2013), popular yet sparsely reviewed products tend to be very desirable for reviewers, as these products maximise their probability of exposure. Marinescu *et al.* (2018) study the implications of reviewing behaviour motivated by nudging and incentives. They observe that the greatest proportion of voluntary reviews is assigned 5* and 1* ratings, with a larger proportion of voluntary reviews assigning a 5* rating. This gives rise to positively skewed rating distribution. The rating distribution of incentivised reviews, however, resembles a somewhat normal distribution. This dissertation hypothesises that voluntary reviews are driven by behavioural biases, whereby people with extreme opinions are more motivated to share them, compared to people with moderate opinions.

2.2 Inference from feedback communities

Hu *et al.* (2008) employ transaction cost economics, developed by Williamson (1979), and uncertainty reduction theories to assess the effectiveness and temporal effects of reviews on Amazon product sales. They found that the impact of online reviews on sales follows a decreasing trend over time, which is also supported by Basuroy *et al.* (2003). Further, Etumnu *et al.* (2019) find that the sampling distribution of ratings, skewness and variance, affects product sales, with higher selling products being most positively affected. This sheds light on the role of disconfirmation and conformity in reviewing behaviour, which is examined in this dissertation, by showing how early reviewers have a more powerful say. Hu *et al.* (2008) also show that consumers pay attention to review attributes besides ratings, such as pictures and helpfulness votes. Shen *et al.* (2013) and Kaushik *et al.* (2018) support this notion by observing that the number of helpfulness votes significantly affects the perceived quality of products and, consequently, product sales. Adomavicius *et al.* (2018) study how recommender systems, which suggest relevant products to customers, affect prospective buyers' willingness to pay (WTP) for specific products. The paper analyses the results of two controlled behavioural studies and concludes that recommendations significantly affect WTP. Identifying and correcting behavioural biases in reviewing behaviour could improve recommender systems by eliminating incorrect suggestions and thus maximising the likelihood of correct mapping between sellers and customers.

2.3 Biases in reviewing behaviour

Hu and Pavlou (2009) point out that due to J-curve rating distribution, the validity of the average as an indication of a product's quality is distorted. They note that having a wrong proxy for product quality gives rise to biased hypothetical quality distribution of products. Consequently, reviews send incorrect signals and recommendations to prospective buyers, causing less optimal purchasing decisions. Chevalier and Mayzlin (2006) note that an overwhelming majority of reviews are positive. Hu and Pavlou (2009) explains these patterns with *purchasing bias*, whereby buyers develop a connection to the purchased product and overstate its value, likely resulting in a 5* rating. Comparing experimental and collected data from Amazon.com, Hu and Pavlou (2009) find that 5* reviews account for 55% of the population, whereas under experimental setting, this proportion was only 8%. This is consistent with Kahneman *et al.* (1991) who show that people tend to overvalue things they own. Hu and Pavlou (2009) also discuss that

under-reporting bias can explain why we observe denser tails in the rating distribution. Intuitively, reviewers will not significantly benefit from posting a neutral review due to a lack of emotional attachment. Positive and negative reviews are often thought to be correlated with higher emotional attachment, which is examined in this dissertation. Powell *et al.* (2017) show that prospective buyers fail to make meaningful inference about product quality due to *popularity bias*. They explain that prospective buyers are attracted by the number of reviews a product has and incorrectly correlate the product's number of reviews with its quality. As Kawamura (2011) notes, reviewers tend to exaggerate their positive or negative experience because of a need for attention, distorting the sampling distribution of ratings.

2.4 Implications of biases and policy recommendations

Che and Hörner (2018) illustrate that recommender systems' ability to suggest new products is limited by lack of reviews due to crowding out from popular products. Their paper proposes to solve this problem by spamming, where the system randomly recommends customers products with no earlier information. This allows random exploration of new products, gathering review information and ensure better future suggestions. Brynjolfsson *et al.* (2006) argue that recommendations help consumers try new products they would not otherwise purchase. Contrarily, Mooney and Roy (2000) argue that recommendations reinforce the sales of already popular products, reducing sales diversity and worsening the rating distribution. Fleder and Hosanagar (2007) tests the effect of recommendations on sales diversity using simulation model with Gini coefficient by Gini (1912) and supports the idea that recommendations decrease sales variety. Shen *et al.* (2013) and Kaushik *et al.* (2018) discuss that prospective buyers highly value the number of helpfulness votes when deciding upon future purchases. Consequently, displaying reviews with more helpfulness votes on the first pages would nudge consumers to make better purchasing decisions. Keser (2003) and Luca (2017) propose creating reputation mechanisms within feedback communities. Such a mechanism includes forums and top-sellers interviews, allowing prospective buyers to have a multi-dimensional opinion of the seller. According to Marinescu *et al.* (2018), providing ethical incentives to encourage engagement in feedback communities can help reduce biases. When more users with mediocre experience review, feedback communities will provide more accurate product quality distributions.

3 Theoretical underpinnings

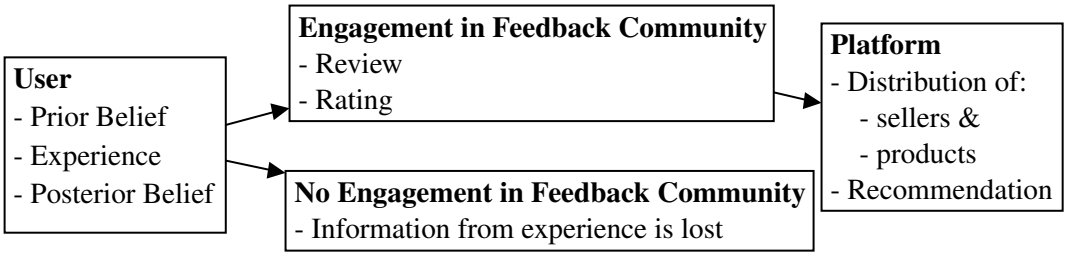
This section outlines the role of feedback communities and their importance to recommender systems. We introduce a set of biases seemingly observed in reviewing behaviour. Lastly, we outline essential statistical theories applied in our empirical analysis.

3.1 Feedback communities and recommender systems

Feedback communities provide a platform where past and prospective buyers meet to exchange information. A feedback community aims to utilise an aggregation of reviews to create a hypothetical distribution that ranks products or services. Prospective buyers then utilise this hypothetical distribution to make a more informed decision on their future purchases, eventually yielding a higher utility. The diagram below illustrates the reviewing interaction between past and potential buyers in feedback communities. Past buyers, *users*, utilise prior belief and experience (likelihood) to form a posterior belief about product quality. They then have a choice whether to review the product or not. Reviews give rise to the distribution of sellers and products utilised by potential buyers when making choices or recommender systems when suggesting products. After potential buyers utilise the *platform* for making a purchase, they become *users* and join the feedback community with the choice to review or not.

O'Donovan and Smyth (2005) highlight the importance of recommender systems in building trust between users and systems, which would allow users to reach higher utility. Feedback communities collect data from past buyers and feed the recommender systems with information, necessary to make informed suggestions to prospective buyers. The collection of data from feedback communities comes in various forms. In Amazon.com, for example, preference revelation arises from what the user purchases, browses or reviews. This dissertation mainly deals with preference revelation mechanisms from reviews. The recommender system receives such information and then recommends products the system expects the user to enjoy.

Diagram 1. Interaction of past buyers and potential buyers in the feedback community



3.2 Behavioural biases

This section discusses essential behavioural biases used when understanding the effects of self-selection and social influence on reviewing behaviour.

3.2.1 Self-selection

As defined in Heckman (1990), self-selection arises when procedures other than simple random sampling are implemented to sample observations from the studied population. Thus, the sample is a manipulated representation of the actual population, influencing inference from the data. The studied sample of reviewers consists of people who decide to post a review and thus self-select to belong in feedback communities. The paper outlines several varieties of self-selection, of which engagement self-selection and purchasing self-selection are discussed throughout this dissertation.

3.2.2 Reciprocity

As discussed in Falk and Fischbacher (2006), reciprocal behaviour arises when people reward actions they perceive as kind and punish unkind ones. The paper further discusses that people take into account the outcome of the action and the intentions. Relating to our analysis, Falk and Fischbacher (2006) also mentions how such behaviour may lead to unfair and inaccurate distributions in marketplaces. Consistent with our observation, rating behaviour is altered by reciprocity, resulting in product ranking inconsistent with our prediction under no reciprocity.

3.2.3 Social influence

Rashotte (2007) defines social influence as an adjustment in peoples' behaviour and attitudes induced by interaction with other people. Regarding online feedback platforms, Sridhar and Srinivasan (2012) specifies that consumers' purchasing choices are affected by current product ratings. Their findings highlight that customers who review are themselves influenced by already existing reviews.

3.2.4 Conformity and anchoring

As investigated in Bernheim (1994), conformity refers to the concept that people are averse to a change of the status quo and prefer to follow the most popular attitude rather than deviating from it, even if this leads to a violation in their utility preferences. Anchoring was first applied to human behaviour by Sherif *et al.* (1958) and these findings were further discussed in Tversky and Kahneman (1974). Anchoring refers to the situation where judgements of different matters are altered because of some initial information that influences consequent decision making. Although we would want reviewers' feedback to reflect their experience, we find evidence that their feedback is affected by previous ratings of past reviewers of the same product. We observe that reviewers are averse to deviate from current product ratings and anchor current ratings to form expectations.

3.3 Ordinal data

In this section, we provide a brief statistical theory on the limitations of ranking ordering used in Amazon.com. Ordinal data, Stevens *et al.* (1946), allows for rank-ordering. Taking the average of ordinal data would not be appropriate in such circumstances since the intervals from different ranks are not defined to have equal width. For instance, in ratings, although the difference between 4 and 5 is equal in magnitude to the difference between 1 and 2, we cannot confidently say that in reviewers' perception, this difference is equal. In the same manner, we cannot confidently say that someone who rates a product with a rating of 5$*$ prefers the product 5 times more than someone who rates the same product 1$*$ and lastly, if someone's experience lies in the middle of two ranks, for example at 2.5, the reviewer is then forced to either inflate or deflate their experience. The above illustrates some basic statistical issues when using the mean to make inferences on products, and Acemoglu *et al.* (2017) show that different feedback communities may provide different conclusions due to the use of ordinal data.

Liddell and Kruschke (2018) perform an in-depth computational analysis on why metric statistics such as the mean should at most cases be avoided when dealing with ordinal data as it can lead to detection of effects where there are none (type I error) and, on the contrary, miss out on causal interpretations (type II error).

4 Data

4.1 Data source

For our analysis, we considered datasets from a range of feedback communities collecting customer reviews. We chose data from Amazon.com for the following reasons. Firstly, Amazon offers review data on a great variety of products organised into 29 categories, and this allows us to make insightful inference on reviewing behaviour across different sub-markets. Secondly, the business model of Amazon is heavily dependent on ratings and reviews as means to bridge the trust between sellers and buyers, which is typically absent in online marketplaces and help users choose between competing products. Further, besides a feedback community, it also represents an online marketplace, where users can not only review but also purchase products. The data in our analysis is collected by Ni and McAuley (2019), and records information from 233.1 million reviews from May 1996 to October 2018 and are available at the following website (maintained by a team at the University of California, San Diego). As Amazon.com no longer publishes review data after October 2018, the used dataset represents the most recent accessible data.

4.2 Data manipulation

The data is available in JSON data format. Given the complex nature of the data structure and vast amounts of observations, it was essential to use sophisticated programming techniques using R to process them into Stata-accessible format. The data is organised into 29 data files according to product categories. To streamline our analysis and allow comparisons between categories, all files were pooled together, creating the original dataset with more than 233 million observations. Given the substantial sample, we decided to reduce our observations to ensure that viable and efficient analysis can be performed while preserving high precision, accuracy and confidence in our results. The individual subsets of the original dataset are listed in the table below.

Table 1: Description of all subsets of original dataset.

Datasets	Number of Reviews	Categories	Sentiment Analysis	Reviewer 5-core	Product 30-core
Original	233,100,000	29	no	no	no
Alpha	72,400,501	16	no	no	no
Beta	8,960,896	16	no	yes	yes
Gamma	614,377	1	yes	yes	yes

Sentiment Analysis: We performed sentiment analysis in this dataset.
Reviewer 5-core: Each reviewer in our dataset has made at least 5 reviews.
Product 30-core: Each product in our dataset has at least 30 reviews.

4.2.1 Alpha

In the first stage, we selected 16 out of 29 product categories, eliminating those categories which were over dominating or dominated in the sample. Specifically, we did not include datasets that were too small and too large in terms of the number of observations. We ensured a balance in our dataset regarding the elasticity of demand by including products categories regarded as luxury goods and categories regarded as necessities. In the second stage, we ensured the data in our sample was correctly set out and matched our source. Our investigation led to the realisation that some reviews were duplicates. We confirmed that concerning observations were exact duplicates by cross-validation with the primary data-source. The Stata built-in command `duplicates` was used to identify such observations and showed 100% match in all the primary variables. This two-stage process resulted in a data-subset with 16 product categories and 72,400,501 reviews. Lastly, we randomly chose several products in our processed dataset and cross-validated the accuracy of the information collected. We compared our data to the published reviews available on Amazon.com and found complete matching of information.

4.2.2 Beta

To improve the credibility of collected data, we further constrain the Alpha dataset by imposing the following restrictions. Only customers who reviewed at least five times and only products which received at least 30 reviews were contained in the dataset. The constraints were selected after careful consideration, ensuring a balance between sample efficiency and diversity. The product constraint ensures that all products in this sample represent frequently bought products that can be regarded as trustworthy. The reviewer constraint is in place to account for fake reviewers. The Beta dataset includes around 8,960,000 million observations.

4.2.3 Gamma

Taking the previously specified dataset Beta as the baseline, we further restrict the dataset by keeping only observations that include information about the price of the reviewed product. Such information is only available for the category of office products. This process further reduces the data dimension to approximately 620,000 observations. Given the complex structure and long data generating process of the sentiment analysis, the algorithm could only be used in a suitably smaller sample. This data scale is thus ideal for performing sentiment analysis on the review text.

4.2.4 Sentiment analysis

Since Amazon does not share personal information about users, our analysis was missing demographic insights. To generate insightful demographic information on the reviewer, we decided to proxy the **gender** variable of reviewers using sentiment analysis on the provided reviewers' name. The implementation of sentiment analysis is detailed in the Appendix. To illustrate this concept, when the program matches "Alex" to gender, it will check how many times "Alex" is associated with each gender. From this, it will calculate the probability of "Alex" being male or female. Although this is the best available method for gender derivation, there are some limitations when extracting the gender from a name. Firstly, some users choose not to provide their real name and choose to provide a nickname from which our package can not confidently identify the gender, producing **unknown** result. Secondly, there are some names which could equally be male or female. Our package deals with this by producing one of three outputs: **mostly male, mostly female** and **andy** (androgynous). The names with a high probability of either gender are classified as **male** and **female**, respectively. To utilise this variable in our regression analysis, we created a **female** binary variable to control for reviewer specific factors.
Sentiment analysis was further applied in the form of text analysis aiding exploration of reviewers' behaviour. We initiated our analysis by extracting the **sentiment score** of the review text. The sentiment score can take negative, neutral or positive sign depending on whether the overall sentiment of the review is negative, neutral or positive. We then extended our analysis to the classification of emotions. We get the number of words that express the following emotions from each review: **anger, anticipation, disgust, fear, joy, sadness, surprise** and **trust**. Moreover, we extract the number of **positive** and **negative** words in each review.

The packages we use in sentiment analysis are discussed in the Appendix, both using the NRC Emotion Lexicon from Mohammad and Turney (2013) to analyse the text from the reviews. The NRC Emotion Lexicon contains a list of words with their respective association to an emotional classification. Various lexicons are available, but the NRC lexicon is one of the most popular since a great advantage of this package is that it can identify connections with preceding and proceeding words.

4.3 Summary statistics

Table 2 provides important summary statistics of the number of reviews, the number of products and the number of reviewers for each of the 16 categories included in the Beta dataset. The majority of the reviews in office products are included in the Gamma dataset. Following the sentiment analysis performed in section 4.2.4, we illustrate primary summary statistics from our exploration in Figures 1 and 2. Figure 1 shows the aggregate number of words associated with each of the measured emotions. It is fascinating to note that reviewers most often use emotionally attached words to trust, illustrating the importance of feedback communities in building trust between sellers and buyers.Figure 2 is a histogram of the sentiment score. We observe an overwhelmingly positively-toned text, which interestingly resembles a normal distribution. Further, the highest frequent score is at 0, which refers to a neutral text. This seems surprising at first sight. Further investigation led to realisation that some reviewers tend to spend little time writing review text, providing review text with only a few words. Therefore, the algorithm does not have enough emotional words to accurately measure the sentiment score and return 0, corresponding to a neutral emotional tone.

Figure 1: Words emotion classification of reviews **(Gamma dataset)**

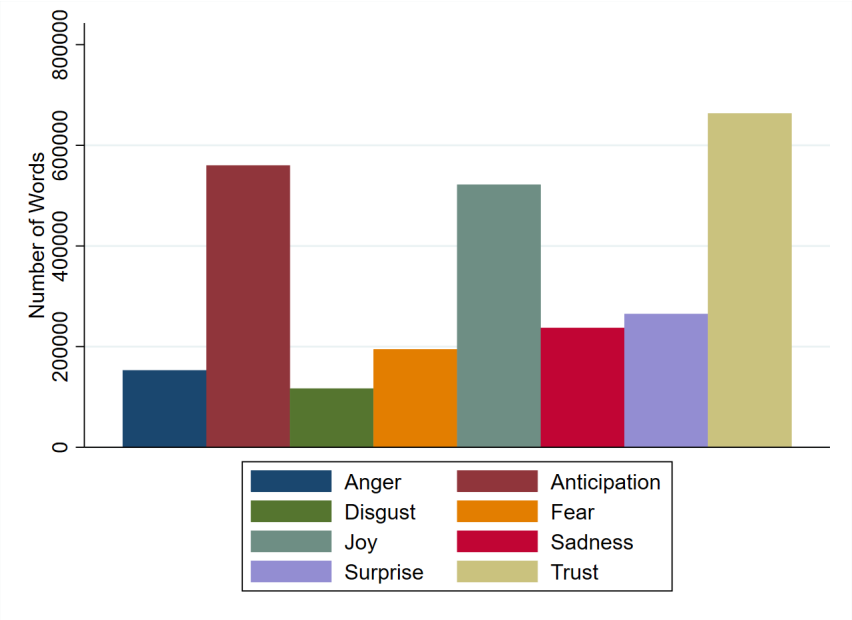


Figure 2: Histogram of sentiment score for reviews **(Gamma dataset)**

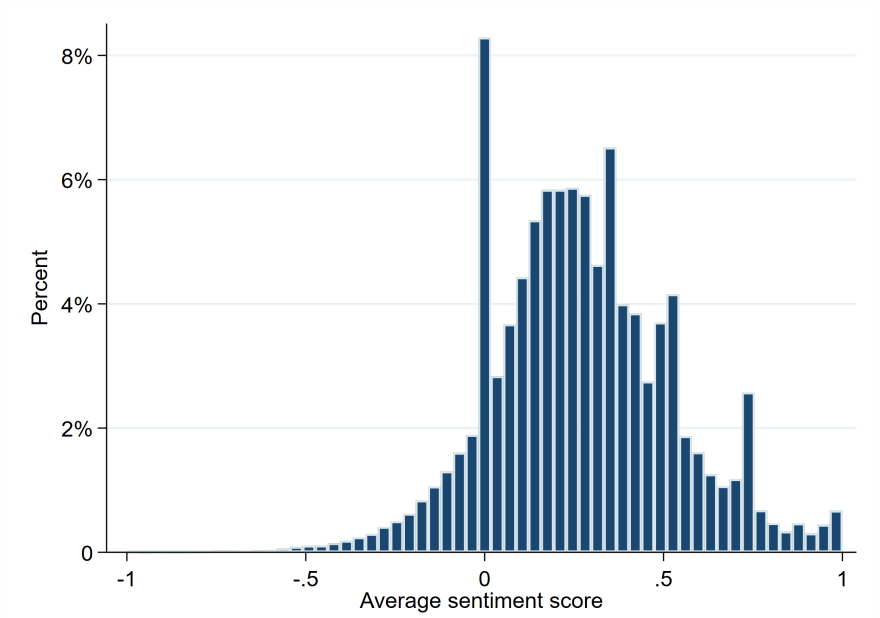


Table 2: Summary statistics **(Beta dataset)**

Product Categories	**Number of Reviews**	**Number of Products**	**Number of Reviewers**
Arts, Crafts and Sewing	201,659	2,838	28,252
Automotive	892,458	10,853	113,440
CDs and Vinyl	731,130	10,031	86,047
Cell Phone and Accesories	688,856	8,188	109,491
Digital Music	35,964	642	5,378
Fashion	10,162	2,082	2,929
Grocery and Gourmet Food	643,505	6,832	70,719
Industrial and Scientific	23,016	389	2,071
Luxury Beauty	8,613	103	856
Musical Instruments	119,250	1,401	16,675
Office Products	769,721	25,855	99,500
Patio lawn and Garden	419,242	5,498	43,634
Pet Supplies	1,472,605	11,181	159,071
Prime Pantry	79,931	1,108	6,067
Sports and Outdoors	1,633,608	18,460	207,519
Tools and Home Improvement	1,231,176	13,959	126,895
Total	**8,960,896**	**119,420**	**1,078,544**

4.4 Description of variables

In this section, we describe the variables used in our analysis. We also explain how we extract and generate further variables to aid our exploration. Our dataset is at the review level, and this implies that every observation in our sample is a different review.

4.4.1 Dependent variables

Our analysis uses two dependent variables which are closely related, **rating** and **extreme rating**. Rating $r \in [1^*, 2^*, 3^*, 4^*, 5^*]$, where $r = 1^*$ being the worst and $r = 5^*$ being the best possible outcome. Extreme rating is defined as a binary variable, which takes a value of 1 when $r = 1^*$ *or* 5^* and a value of 0 when $r = 2^*$ *or* 3^* *or* 4^*. Testing our model on two closely related dependent variables provides robustness and confidence in our results and helps discover rating behaviour more closely.

4.4.2 Primary independent variables

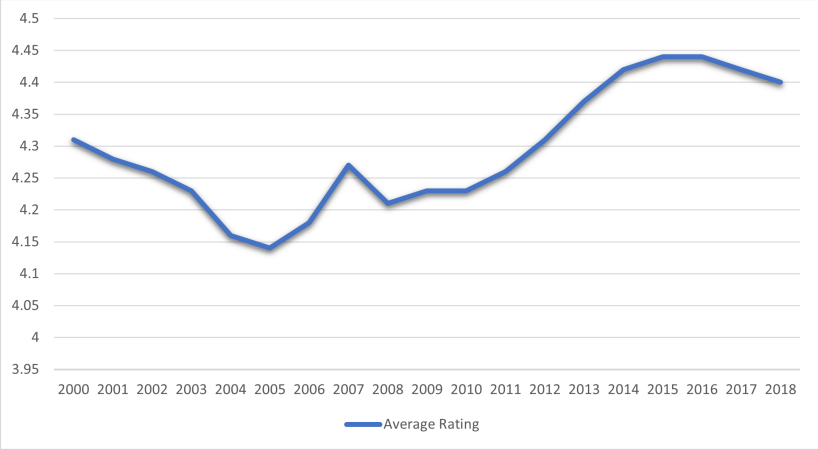
Our main variable of interest is **disconfirmation** (δ). We define disconfirmation as the difference between actual quality and expected quality. Although it is non-trivial to measure such variables, we follow a simplified version of the modelling

used in Ho *et al.* (2017). Expected quality is essentially the expectations reviewers set before purchasing a product. Consider a review on product x from a reviewer at time t. We measure expected quality as the average rating on product x from time 0 to time $t-1$, where time is measured daily in our dataset. Ideally, we would like to measure the average rating before the purchase date, but our data does not provide such information, so we use the day of the review instead. The actual quality is the quality that the reviewer perceives on the product post-purchase. We assume this to be uniform across reviewers on the same product, and we use the overall average rating of each product as actual quality. Ultimately, if the reviewer perceives the product as better than expected $\delta > 0$ and if the reviewer perceives the product as worse than expected $\delta < 0$. To account for the possibility of a non-linear trend (concave or convex), we also consider δ^2 and δ^3. Moreover, an important aspect of this paper relies on sentiment analysis of the review, including emotion classification and **sentiment score** as discussed in sentiment analysis. We also utilise the absolute value of the sentiment score when running the main aspects of our analysis. Powell *et al.* (2017) discusses social influence in feedback communities by examining how reviewers' rating is affected by past users, both in ratings and number of reviews. Therefore, we include the **expected quality** and the number of past reviews in our model to observe those effects and compare them with prior literature. Lastly, to aid our analysis of self-selection reviewing behaviour, we also calculated the **average extreme rating** and the **number of reviews per reviewers**.

4.4.3 Control variables

We tested the performance of our models with different sets of control variables and chose the ones giving the best-fitted results. We also included yearly dummies to control for the effect of years on rating behaviour. Figure 3 shows the trend of average rating across all our categories over 18 years, where we have information on plenty of categories. When running regressions using the Beta dataset, we also include product **category** dummies. Although we are not explicitly interested in the effect of our control variables on rating and extreme rating, some control variables yield results that could shed light on our primary analysis.

Figure 3: Yearly trend for average rating across all product categories **(Beta dataset)**



On the review level we are provided with the number of **helpfulness votes** the review receives, the **review text**, the respective **summary** and the binary variable **verified**. Verified purchase means that Amazon can guarantee that the reviewer has confidently purchased the product from Amazon.com and did not receive the product at a deep discount. Deep discounts used by some sellers to try and incentivise positive ratings is a practice that can jeopardise the feedback community as the review will not be a fair mapping from the user's experience.
On the product level, we have the unique product id, "Amazon Standard Identification Number" (**Asin**). From the meta datasets available to us, we were able to extract the **price** of some products. We have information on the unique **reviewer id** and the reviewer's name on the reviewer level.
To develop models that we will utilise in our analysis, we define the following variables we generated. Firstly, we generated a **word count** of the text of both the **review** and the **summary** of the review. This helps us capture the "cost" of writing a review in terms of time spent. We adopt a reasonable assumption that the greater the amount of words in a text, the more time is spent writing that review. We then manipulated the rating of each review along with various other parameters to calculate the **average ratings** per **reviewer**, **category**, **product** and **year**.

5 Empirical specification

Besides the graphical representations of data, which help answer a range of questions, our empirical study relies on appropriate regression analysis using logit and ologit models. In this section, we discuss and justify our methodology by considering statistical assumptions and alternative models.

5.1 Choice of regression models

We use the logistic, Wright (1995), and ordered logistic model, Grilli and Rampichini (2014), for extreme rating and rating to run our analysis. Logistic models are functioning non-linearly and use the logistic distribution to analyse the relationship of primary variables and covariates on the dependent variable.

$$G(z) = \Lambda(z) = \frac{exp(z)}{1 + exp(z)} \tag{1}$$

Limited dependent variable models such as logit and ologit use maximum likelihood estimation to estimate coefficients. The optimum is found using iterative processes until convergence is found. Our models reached convergence, which increases confidence in our method against the existence of quasi-complete separation patterns, Allison (2008). The maximum likelihood estimator maximises the likelihood of observing the given sample from a population. This is the value that maximises the likelihood equation of our regression model and gives us the coefficients of our analysis. The likelihood function, which we aim to maximise, follows the following equation, Greene (2000).

$$\begin{aligned} \hat{\beta} &= \arg\max_{\beta} \left[\ln \mathcal{L}(\beta)\right] \\ &= \arg\max_{\beta} \left[\sum_{t} \left(y_t \ln \left(\frac{\exp\left(x_t'\beta\right)}{1 + \exp\left(x_t'\beta\right)} \right) + (1 - y_t) \ln \left(\frac{1}{1 + \exp\left(x_t'\beta\right)} \right) \right) \right] \end{aligned} \tag{2}$$

The results presented in tables 3 and 4 represent the average marginal effects, meaning that the effects are calculated for each observation and then averaged. Intuitively, it is the average change in the probability of posting, say 1* review, for a unit change in disconfirmation. Extending the empirical analysis, we run a series of logit regressions which take the extreme rating as the dependent binary variable. Logistic interpretation is shown in equation 2. Y represents our dependent variable, and **X** our matrix of explanatory variables.

$$\mathbb{P}(Y) = \frac{exp(\beta_0 + \boldsymbol{\beta X})}{1 + exp(\beta_0 + \boldsymbol{\beta X})} \tag{3}$$

5.1.1 Model assumptions

The main assumptions which need to be taken into account when implementing logit and ologit regressions are the independence of errors, absence of multicollinearity, omission of influential outliers and model exogeneity. We account for the extreme observations by restricting our sample to products with at least 30 reviews and reviewers who posted at least 5 reviews. This helps ensure a less biased sample, as we will discuss in later sections. Further, to control for endogeneity, we account for product, reviewer and yearly effects with appropriate control variables to validate the independence of errors assumption. Additionally, we use cluster robust standard errors to account for heteroscedasticity across our clusters. To perform ologit regression, we had to ensure that the proportional odds assumption holds, Peterson and Harrell Jr (1990). This means that regardless of which partition among the ordering scale we considered, the slope approximation between pairs is equal across all rating levels. Lastly, the statistical software can identify the possible multicollinearity and automatically ensure this assumption when running regressions. As we discuss in limitations and given that every reviewer has a different personality and judgement criteria, it is impossible to model a perfectly exogenous regression. Our data set does not provide enough information regarding the reviewer to control for the above assumptions. Modelling reviewers' personality characteristics is crucial in understanding behavioural biases and should not be omitted by empirical techniques. Our results must be interpreted carefully, provided that empirical assumptions are not fully satisfied. We, therefore, cannot make a fully confident causal inference from our results. Nonetheless, the model specified serves as the best possible representation, given the data limitations.

5.2 Model specification

We investigate the effect of disconfirmation and expected quality on the probability of observing a specific rating on the scale from 1* to 5* and the probability of observing an extreme rating corresponding to a 1* or 5* rating. Upon various considerations, we present and discuss the results from the following models.

$$\begin{aligned}\mathbb{P}(rating\ /\ extreme\ rating) = {} & \beta_0 + \beta_1 * absolute\ disconfirmation + \\ & \beta_2 * |disconfirmation|^2 + \beta_3 * |disconfirmation|^3 + \\ & \beta_4 * expected\_quality + \beta_5 * yearly\ dummies + \epsilon\end{aligned}$$

(Model 1)

$$\begin{aligned}\mathbb{P}(rating\ /\ extreme\ rating) = {} & \beta_0 + \beta_1 * absolute\ disconfirmation + \\ & \beta_2 * |disconfirmation|^2 + \beta_3 * |disconfirmation|^3 + \\ & \beta_4 * expected\_quality + \beta_5 * yearly\ dummies + \\ & \beta_6 * number\ of\ past\ reviews + \beta_7 * verified + \\ & \beta_8 * word\ count\ of\ review + \\ & \beta_9 * word\ count\ of\ summary\ review + \\ & \beta_{10} * number\ of\ product\ reviews + \\ & \beta_{11} * number\ of\ reviewer\ reviews + \epsilon\end{aligned}$$

(Model 2)

$$\begin{aligned}\mathbb{P}(rating\ /\ extreme\ rating) = {} & \beta_0 + \beta_1 * absolute\ disconfirmation + \\ & \beta_2 * |disconfirmation|^2 + \beta_3 * |disconfirmation|^3 + \\ & \beta_4 * expected\_quality + \beta_5 * yearly\ dummies + \\ & \beta_6 * number\ of\ past\ reviews + \beta_7 * verified + \\ & \beta_8 * word\ count\ of\ review + \\ & \beta_9 * word\ count\ of\ summary\ review + \\ & \beta_{10} * number\ of\ product\ reviews + \\ & \beta_{11} * number\ of\ reviewer\ reviews + \\ & \beta_{12} * vote + \beta_{13} * female + \beta_{14} * price + \\ & \beta_{15} * female\ x\ verified + \\ & \beta_{16} * absolute\ sentiment\ score + \epsilon\end{aligned}$$

(Model 3)

5.2.1 Alternative models

This section justifies the suitability of our model specification by discussing the alternatives. Initially, we attempted to use the linear probability model; however, upon observing a statistically significant effect on the higher power coefficients of disconfirmation, we decided to use the logistic model instead. Further, unlike mlogit, in which the dependent variable takes several values, which are cardinal, in ologit the higher value of the variable corresponds to a higher value of the rating satisfying the ordering principle of the dependent variable in our study. Lastly, we cross-validate our results using probit and oprobit, respectively. We conclude that the effect differentials are negligible, and both models are suitable. This idea is supported in various academic papers such as Allison (1999). Although, as we show in robustness check, the differences are negligible, we conduct our analysis using logistic models due to its nature of longer tails than probit and a likelihood value for our data.

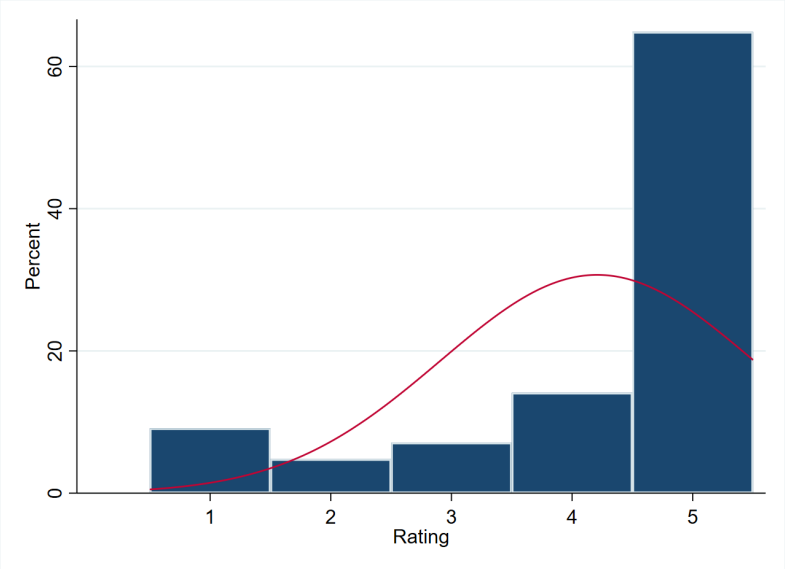
6 Results

This section analyses the sampling distribution of review ratings and observes that the distribution is asymmetrically bimodal, resembling a J-curve shape. Further, we identify and examine two behavioural biases in reviewing behaviour that could drive this distribution. Specifically, we investigate the presence of engagement self-selection bias via reciprocity and the presence of social influence via disconfirmation.

6.1 Sampling distribution of ratings on Amazon.com

In this subsection, we test the sampling distribution of ratings using the Alpha dataset. Initially, we plot a histogram of ratings in our dataset and observe the respective densities in percentiles. Figure 4 depicts the distribution together with the appropriately scaled normal density line shown in red. The density line demonstrates how our data would look if its distribution were normal. We identify that the observed distribution is skewed and not-normal compared to the scaled normal density line. The high end and the low end of the distribution are disproportionately larger than expected. On the contrary, the middle values are lower than expected.

Figure 4: Sampling distribution of ratings.**(Alpha dataset)**



The distribution we observe is J-curve shaped, violating the distributional assumption of mean inference. The middle of the distribution is less dense than the tails, with the high-end tail being much denser than the low-end tail. Since ratings are categorical, the most appropriate test of approximate normality for discrete data is Pearson's chi-squared test, Plackett (1983). To perform this test, we specify the proportions we expect to observe under the null hypothesis. The null hypothesis defines that the sample comes from an approximately normal distribution with the following proportions for the five ratings: 10% of 1*, 20% of 2*, 40% of 3*, 20% of 4* and 10% of 5* reviews. The test uses 4 degrees of freedom and yields a huge test statistic. We can confidently reject the null hypothesis at all significant conventional values in favour of the alternative, namely that the distribution does not follow approximate normality. Even though our data is not continuous, the large sample allows us to perform Quantile-Quantile (Q-Q) plot and run the Kolmogorov–Smirnov test, Kolmogorov (1933). The chi-squared probability plot and the Q-Q plot are shown below. We reach the same conclusion that our data does not come from an approximately normal distribution.

Figure 5: Chi-squared probability plot (4 d.f.), for the sampling distribution of ratings. **(Alpha dataset)**

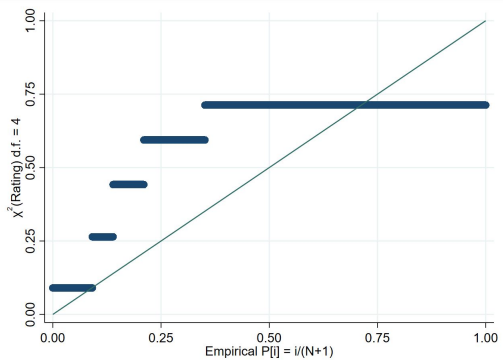
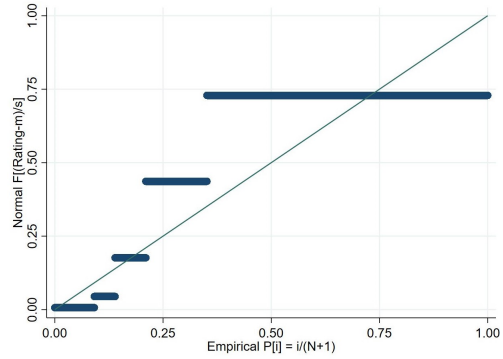


Figure 6: Normal quantile-quantile probability plot for the sampling distribution of ratings.**(Alpha dataset)**



Hu and Pavlou (2009), Hu Nan (2006), Marinescu *et al.* (2018) and Dellarocas *et al.* (2006) use various empirical tests to test both the parent sampling distribution and the approximate distribution of ratings as the number of included reviews becomes larger. Our results are consistent with their findings.

Lastly, to give robustness to our results, we checked some summary statistics from the sample data that make up this distribution. The median and mode value is at 5*, and the mean is at 4.21*. Skewness is a measure of symmetry, and a value of -1.52 implies a rightly skewed distribution. Kurtosis tries to measure how large the asymmetry is, and a value of 3.93 implies a very high-tailed and peaked distribution. Both measures of fit support the presence of an asymmetrically bimodal distribution, with skewed tails and a denser high-end tail.

6.2 Engagement self-selection bias via reciprocity

Engagement self-selection bias refers to the observation that not all types of customers engage in feedback communities equally often. We identified customers who never review, reviewers who contribute regularly (frequent) and the occasional reviewers (infrequent).

6.2.1 Empirical analysis

By performing empirical analysis on the Beta dataset, we test the hypothesis that the infrequent reviewers are motivated by reciprocity and post a review to either reward the seller with a 5* review or punish them with a 1* review. Figure 7 demonstrates how the average rating of a reviewer varies with the number of reviews the reviewer posts. We observe a positive trend whereby the average rating is higher for the reviewers who post many reviews. Note that the scatter plot does not examine a dynamic process. Instead, a single observation on the scatter plot constitutes a single reviewer whose average rating is calculated based on all reviews he posted up to October 2018. Examining how rating behaviour changes with review frequency, Figure 8 shows how the proportion of extreme ratings varies with the number of reviews the reviewer posts. Note that the average extreme rating per reviewer corresponds to the proportion of extreme ratings, i.e. either 1* or 5*. For illustration, consider a reviewer who posted the following reviews; 5*, 4*, 4*, 1*, 5* and 2*. Three of the posted reviews are extreme, resulting in the average extreme rating of this reviewer being 3/6 = 0.5. We observe that infrequent reviewers are, on average, more likely to post a review with an extreme rating. Further, we observe that more frequent reviewers tend to rate more often 2*,3* and 4* than infrequent. Intuitively, although frequent reviewers post regularly and under various conditions, infrequent will post a review under certain conditions, such as reciprocity.
These results are also supported by the OLS regression performed on the fitted values in Figure 8, giving a negative and statistically significant coefficient. Note that most values lie outside the confidence interval for the fitted line. This is justified, given the incredibly large sample. This regression takes into account possible heteroskedasticity with cluster robust standard errors. Results in Table 3 support our methodology by giving a statistically significant coefficient on the number of reviews per reviewer regressed on extreme rating. In context, our result illustrates that the probability of a reviewer publishing an extreme rating falls by 5% for each additional 100 reviews she contributes to the feedback community.

Figure 7: Scatter plot showing the average rating of each reviewer against the number of reviews each reviewer has. **(Beta dataset)**

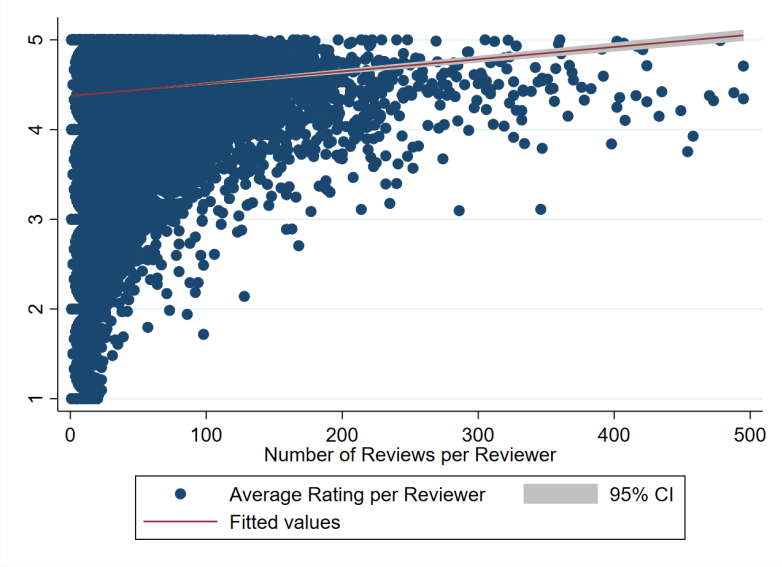
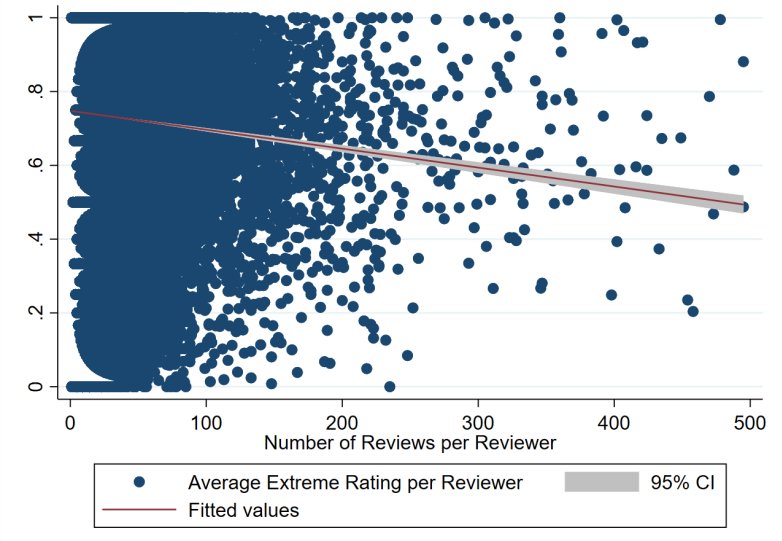


Figure 8: Scatter plot showing the average *extreme* rating of each reviewer against the number of reviews each reviewer has.**(Beta dataset)**



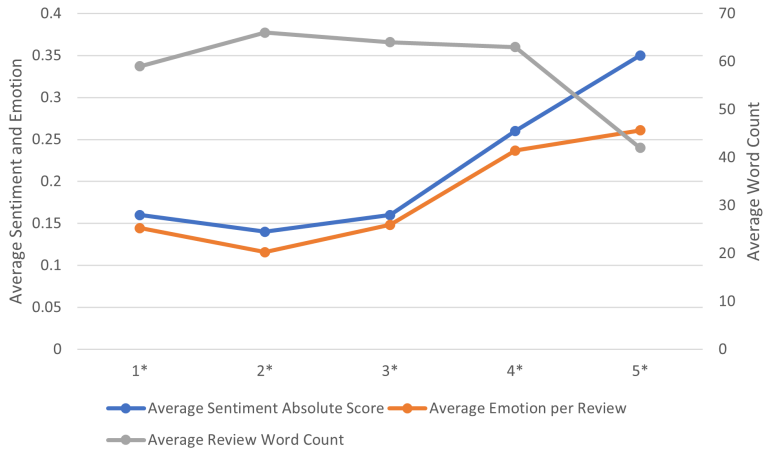
The above results reveal an unusual observation whereby the average rating and the proportion of extreme ratings per reviewer vary with experience, which is inconsistent with the rational view. We speculate that the presence of engagement self-selection bias can explain this observation. Indeed the larger proportion of extreme ratings among the infrequent reviewers seems consistent with our hypothesis that they are likely motivated by reciprocity. Specifically, infrequent customers post reviews to reward the seller with a 5* review or punish them with a 1* review. As expected, our analysis has inconsistent variance across the level of the number of reviews per reviewer.

6.2.2 Sentiment analysis

Further, we perform sentiment analysis on the review text on the Gamma dataset to examine how the absolute value of the sentiment score varies with rating. We have used the absolute value of the sentiment score in our analysis because our primary focus is to examine the effect of emotions on reviewing behaviour, irrespective of whether those emotions are positive or negative. Research could utilise our findings and extend this work by distinguishing between positive and negative tone. We observe that reviews with extreme rating have an average absolute sentiment score of 0.34, while reviews with the middle scale rating have an average absolute sentiment score of 0.22. An increasing value of average absolute sentiment score represents an increasingly emotionally toned review text without distinguishing whether the emotional tone is negative or positive.
Analysis of extreme rating behaviour shows that positive reviews are highly more emotional in positive review content than negative reviews in negative review content. To quantify the difference, we observe that a positive sentiment score (0.35) in 5* reviews is more than twice as emotional as a negative sentiment score (0.16) in 1* reviews. The logistic regression model captures the effect of emotions on extreme ratings. Results presented in Table 3 suggest that reviews with higher emotional toned text have a higher probability of being rated by an extreme rating. Referring to the average marginal effects, a 1 unit increase in the absolute value of the sentiment score leads to, on average, 30% increase in the probability of an extreme rating.
To acquire a greater understanding of the relationship between ratings, emotions and word length, Figure 9 illustrates how these variables behave across ratings. It is important to note that the line connecting the dots in the figure is the only representative, and no relationship should be inferred from it. We observe that the sampling distribution of the average absolute sentiment score and the average

emotion per review closely resembles a J-shaped distribution. Consistency between results and the observed sampling distribution of ratings provides plausible evidence for a relationship between emotions and review behaviour. Conversely, the average review word count instead resembles an inverse J curve distribution with lower word count observed for extreme ratings. We speculate that extreme reviews are less comprehensive and thus are less costly for the reviewer to publish. This is a possible mechanism that can explain the higher than expected distribution among the extreme ratings.

Figure 9: The average sentiment absolute score, emotion proportion and average review word count in each rating. The lines connecting the dots are only representative, not causal. **(Gamma dataset)**



Given the larger absolute sentiment scores for reviews with an extreme rating and given the fact that infrequent reviewers are more likely to post a review with extreme rating rather than a middle score, we find plausible evidence that infrequent reviewers are on average more likely to post a review driven by emotional motives. Consequently, these results provide suggestive evidence that the greater emotional resonance among infrequent reviewers could be described by reciprocity. Emotional stimuli largely drive rewarding and punishment behaviour. Extending our analysis, we observe that frequent reviewers tend to write overwhelmingly longer reviews and use transitional words more often than infrequent reviewers. Performing correlation on emotions and helpfulness we find that less emotional reviews are regarded as more helpful, likely because of a more objective content.

6.2.3 Remarks

Our findings are consistent with Ullah *et al.* (2016). Their paper investigated sentiment analysis on various categories on Amazon.com and found that extreme reviews share more emotional content than non-extreme ones and that the positive extreme is highly more emotional than the negative extreme. Our results' consistency with academic literature gives confidence in our results. Interpreting the above results, we speculate that the J-curve shaped distribution is greatly driven by the infrequent reviewers, who post a large proportion of their reviews with an extreme rating and who account for the bulk of reviewers. Since infrequent reviewers review a lot more on the extremes, this boosts both tails of our sampling distribution. Moreover, we show that the skewed distribution comes from a tendency to rate highly, supported by the emotional content in reviews with a high rating. Thus, if feedback platforms included a larger proportion of frequent users, the distribution of the ratings is more likely to resemble a normal distribution.

Table 3: Logistic regression (logit) of *extreme ratings* on *primary variables*, *yearly dummies & extensive set of control variables*. **(Gamma Dataset)**

	(1) Model 1	(2) Model 2	(3) Model 3
Absolute Disconfirmation	0.0183	0.0542***	0.0466***
	(0.00989)	(0.00948)	(0.0127)
Absolute Disconfirmation^2	0.0382**	-0.000236	0.00254
	(0.0128)	(0.0124)	(0.0170)
Absolute Disconfirmation^3	0.0181***	0.0262***	0.0246***
	(0.00350)	(0.00342)	(0.00481)
Expected Quality	0.232***	0.220***	0.200***
	(0.00329)	(0.00326)	(0.00341)
Absolute Sentiment Score			0.295***
			(0.00469)
Number of Reviews per Reviewer		-0.000181***	-0.000568***
		(0.0000500)	(0.0000868)
Yearly Dummies	-	-	-
	(.)	(.)	(.)
Number of Past Reviews		-0.00000458	0.00000729
		(0.00000513)	(0.00000744)
Word Count of Review		-0.000401***	-0.000232***
		(0.0000127)	(0.0000160)
Word Count of Summary Review		-0.00828***	-0.00583***
		(0.000154)	(0.000225)
Verified		0.0241***	0.0209***
		(0.00231)	(0.00330)
Number of Reviews per Product		0.0000112*	0.00000743
		(0.00000468)	(0.00000614)
Helpfulness Votes			-0.00000925
			(0.00000793)
Female			0.0209***
			(0.00169)
Price			0.000127***
			(0.0000258)
Observations	614377	614377	250946

Marginal effects; Standard errors in parentheses

(d) for discrete change of dummy variable from 0 to 1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Ordered logistic regression (ologit) of *ratings* on *primary variables*, *yearly dummies* & *extensive set of control variables*. **(Gamma Dataset)**

	(1) Rating 1	(2) Rating 2	(3) Rating 3	(4) Rating 4	(5) Rating 5
Absolute Disconfirmation	-0.00680**	-0.00483**	-0.00928**	-0.0161**	0.0370**
	(0.00207)	(0.00147)	(0.00283)	(0.00492)	(0.0113)
Absolute Disconfirmation^2	-0.0207***	-0.0147***	-0.0282***	-0.0490***	0.113***
	(0.00261)	(0.00186)	(0.00354)	(0.00612)	(0.0141)
Absolute Disconfirmation^3	0.00641***	0.00455***	0.00873***	0.0152***	-0.0349***
	(0.000745)	(0.000531)	(0.00101)	(0.00175)	(0.00402)
Expected Quality	-0.0592***	-0.0420***	-0.0808***	-0.140***	0.323***
	(0.000806)	(0.000590)	(0.000943)	(0.00131)	(0.00262)
Absolute Sentiment Score	-0.0739***	-0.0524***	-0.101***	-0.175***	0.402***
	(0.00129)	(0.000961)	(0.00149)	(0.00208)	(0.00493)
Number of Reviews per Reviewer	-0.0000967***	-0.0000687***	-0.000132***	-0.000229***	0.000527***
	(0.0000185)	(0.0000131)	(0.0000251)	(0.0000435)	(0.000100)
Yearly Dummies	-	-	-	-	-
	(.)	(.)	(.)	(.)	(.)
Number of Past Reviews	-0.00000113	-0.000000805	-0.00000155	-0.00000269	0.00000617
	(0.00000145)	(0.00000103)	(0.00000198)	(0.00000345)	(0.00000792)
Word Count of Review	0.0000145***	0.0000103***	0.0000198***	0.0000344***	-0.0000789***
	(0.00000228)	(0.00000162)	(0.00000311)	(0.00000538)	(0.0000124)
Word Count of Summary Review	0.00131***	0.000931***	0.00179***	0.00311***	-0.00714***
	(0.0000428)	(0.0000303)	(0.0000567)	(0.0000946)	(0.000217)
Number of Reviews per Product	0.000000519	0.000000368	0.000000708	0.00000123	-0.00000283
	(0.000000882)	(0.000000627)	(0.00000120)	(0.00000209)	(0.00000480)
Helpfulness Votes	0.0000356***	0.0000253***	0.0000486***	0.0000845***	-0.000194***
	(0.00000149)	(0.00000103)	(0.00000192)	(0.00000334)	(0.00000762)
Female	-0.00393***	-0.00276***	-0.00528***	-0.00912***	0.0211***
	(0.000300)	(0.000213)	(0.000411)	(0.000724)	(0.00164)
Verified	-0.00570***	-0.00399***	-0.00753***	-0.0126***	0.0298***
	(0.000605)	(0.000419)	(0.000775)	(0.00127)	(0.00306)
Price	-0.0000113*	-0.00000805*	-0.0000155*	-0.0000269*	0.0000617*
	(0.00000443)	(0.00000315)	(0.00000605)	(0.0000105)	(0.0000241)
Observations	250946	250946	250946	250946	250946

Marginal effects; Standard errors in parentheses

(d) for discrete change of dummy variable from 0 to 1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.3 Social influence via disconfirmation

In this section, we test the hypothesis that social influence affects reviewers via the magnitude of quality disconfirmation by performing a set of logistic and ordered logistic regression models specified in section 5, on the Gamma dataset. For completeness, we performed those regressions on the Beta dataset and further discussed in robustness checks.

6.3.1 Empirical analysis: *extreme ratings*

We study how the probability of posting an extreme rating is determined by the difference between the expected and actual product quality; and the expected quality itself. The variable of extreme rating takes a value of 1 if the review is either 1* or 5* and takes 0 otherwise. We further include a set of control variables and yearly dummies and conclude that the effects do not vary significantly across different model dimensions. Upon observing the pseudo R^2, we find that the predictive power of our model improves by including more control variables. We, therefore, refer to model 3 when reporting quantitative results. For more insights about the robustness of our results, see robustness checks. Note that the number of observations across our regressions is smaller than the total number of available observations. This is because the variable of quality disconfirmation omits the first review of each product as it acts as the benchmark for calculating the expected quality. The first reviewer does not observe an expected quality. We use the absolute value of quality disconfirmation, which symmetrically represents a deviation of the actual quality from reviewers expectations.
Analysing our results presented in Table 3, we observe a positive coefficient on the absolute value of quality disconfirmation. Intuitively, we find plausible evidence that with an increasing divergence between the expected and actual product quality, the probability of an average reviewer posting an extreme rating increases by approximately 4%. We speculate that the reviewer's prior beliefs about the product quality influence the perceived experience. This, in turn, violates the independence axiom. This prior belief, in turn, is affected by previous reviews of the same product. Ideally, we seek reviews to represent the true quality of the product. Still, due to social influence, we find reviews being potentially influenced by the opinion of previous reviewers.
The first few reviewers seem to set the reference point for long term average rating. Moreover, this potentially creates distorted incentives for sellers to offer better product quality to boost positive reviews, followed by a decrease in quality. This

is consistent with findings of prior literature on social influence, which conclude that early reviewers determine product rating, Etumnu *et al.* (2019). Theoretically, this can be explained by conformity bias and specifically the Bandwagon Effect. Upon observing previous reviews, reviewers tend to follow the status quo and post a review that aligns with the average rating due to deviation aversion. Further, the behaviour is determined by anchoring, whereby future reviewers seem to anchor their rating on the existing rating of the product.

6.3.2 Empirical analysis: *ratings*

We further examine how the probability of posting a specific rating is determined by quality disconfirmation and expected quality with the aid of ologit regressions. The average marginal effects of the model with the best fit (equivalent to model 3 in logistic regression) are presented in Table 4. Observing the coefficients on quality disconfirmation, we identify negative coefficients for a rating between 1* and 4* and a positive coefficient for a 5* rating. Intuitively, as the deviation between the expected and actual quality increases, the reviewer is increasingly likely to post a 5* review and increasingly unlikely to give a rating between 1* and 4*. For example, if the deviation between the actual and expected quality is 1, ($|\delta| = 1$), as measured by quality disconfirmation, the average reviewer is almost 3% more likely to post a 5* rating. Further, for the same absolute value of quality disconfirmation, the average reviewer is 1% less likely to rate product with 4* than an almost 0.5% decrease in the probability of 1*. These effects may help explain the tendency of reviewers to rate 5* ratings, as shown in the histogram of the sampling distribution of ratings. However, we should be cautious when interpreting these coefficients, as reverse causality may exist, and the nature of an abundance of 5* ratings may be driving these coefficients in the first place.
The coefficient on $disconfirmation^2$ explains how the change in probability of posting a certain rating changes for a given change in the disconfirmation. We observe that the probability function has an increasing slope for 5* ratings. For illustration, consider Alex received a parcel with a new toothbrush and observe that the quality is almost as expected. Still, the box is a little damaged, which results in a quality disconfirmation of 1. Moreover, Bob received a toothbrush that is not as sturdy as he expected, setting his quality disconfirmation to a value of 3. He further realises that the packaging is damaged, which further increases disconfirmation by 1 unit to a value of 4. The model predicts that the packaging damage will decrease the probability of posting a rating between 1* and 4* more

than proportionately for Alex. Still, for Bob, the damaged packaging will result in a smaller than proportionate decrease in the probability of posting a rating between 1* and 4*. Interestingly, $disconfirmation^2$ is significant for the ologit regression; however, it is not significant for the logit regression. We interpret this coefficient cautiously.

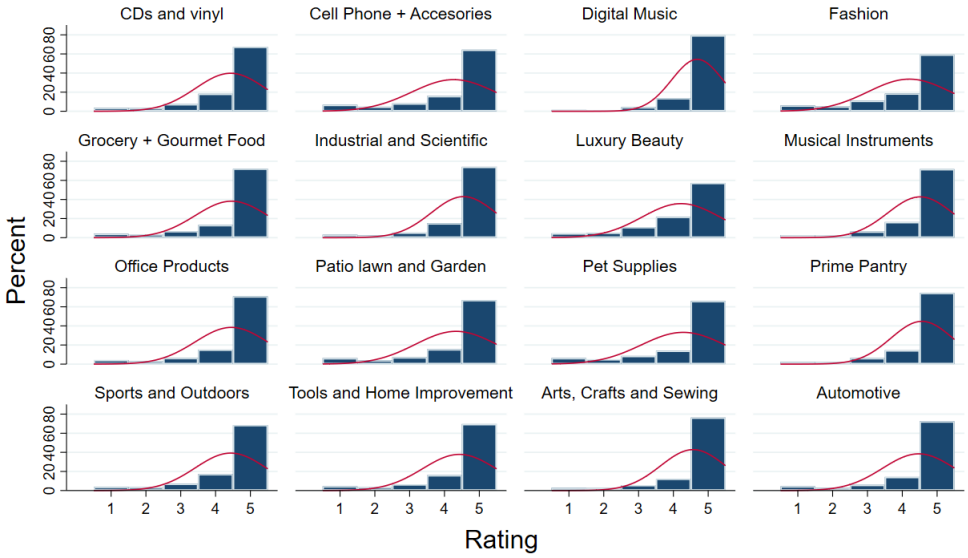
6.3.3 Remarks

Ho *et al.* (2017), Li *et al.* (2020) and Oliver (1977) found evidence that users who experience a higher absolute quality disconfirmation are more likely and quicker to post a review than users with lower absolute quality disconfirmation, and the effects of disconfirmation on product post-evaluation are significant. Our research adds to this literature by finding evidence that the reviewers who experience a higher quality disconfirmation are more likely to review an extreme rating. Further, Ho *et al.* (2017) also indicates that reviewers with higher quality disconfirmation are more likely to post a review that does not resemble their true experience. This finding is consistent with our results of the effects of social influence on reviewing behaviour. Thus, conformity could potentially affect reviewer's rating behaviour and drive ratings to the extremes, explaining the nature of the J-curve shaped distribution. In the absence of social influence, we would expect to see more non-extreme ratings creating a distribution that resembles a normal distribution.

6.4 Robustness checks

We perform the Kruskal-Wallis test, McKight and Najab (2010), for the difference in the average rating between the 16 product categories. Given the distribution of our sample, this test is superior to the analogous ANOVA test because it is a non-parametric test and does not require normality assumptions, Hecke (2012). The test yields a chi-squared statistic of 15 with 15 *d.f.* and a p-value of 0.4514. We thus fail to reject the null hypothesis that the average rating per category differs. Additionally, we observed the sampling distribution of rating across all categories with the aid of histograms and failed to identify observable differences between categories. The empirical analysis utilised the Gamma dataset, which only includes office products. However, given the similar sampling distribution, our analysis would likely yield similar results on data subsets with different category.

Figure 10: Sampling distribution of ratings over categories. **(Beta dataset)**



Further, we tested our results on other product categories and on different datasets, and the analysis yielded similar results. It, therefore, gives confidence that using Gamma dataset containing only one category was robust for making inferences.
We also examine the scatter plots used in our analysis using both the Alpha and Gamma datasets and found consistent results, shown in the appendix. It is important to stress out that when comparing the effects across datasets, we observe a slight deviations in the effects of all regressors on the dependent variables. This is expected due to the differences in sample sizes and control variables. For example, we have no information on price and sentiment analysis on any dataset besides Gamma.
Moreover, we assess whether the results of the regressions differ when using probit and oprobit regression models instead. The regression output is provided in the appendix. Note that we do not show the complete regression output, focusing on the primary variables instead. We fail to identify statistically significant deviation of results from the used logit and ologit regressions. As noted before, we run our analysis on two dependent variables to get a broader picture of rating and reviewing behaviour. The similarity of the results on both regressions shows the consistency in our findings.
Lastly, we attempted to perform conventional goodness of fit tests on our empirical models. We observe that the p-value converges to zero rather quickly, which could imply a bad fit. However, we should be cautious when interpreting the p-value and

the extremely high t-statistic due to them, being affected by our extremely large sample size, as shown in table 1, because the test statistic is an increasing function of sample size. Large sample although, benefit our research in terms of high precision and power. The case for the above argument is supported in Lin *et al.* (2013) and Faber and Fonseca (2014). Thus, to test our model's fit, we first tested our empirical model on smaller subsets of our data, which yielded results that could exhibit confidence. We then performed scalar measures of fit such as pseudo R^2 and information measure tests such as BIC and AIC (Kuha (2004)). Though we run and observe both BIC and AIC, we focus more on BIC due to its bayesian nature and because it deals better with false positives than AIC. As suggested by Ifrach *et al.* (2019) and Acemoglu *et al.* (2017), BIC is suitable given the fact that reviewers are thought to act as bayesian learners and false positives are a greater worry for our modelling robustness. Our paper includes the empirical models with the best measure of fit results, such as the highest McFadden's R^2 and adjusted R^2, McFadden (1974).

7 Conclusion

As discussed in Kawamura (2013), summarizing an aggregation of information is a challenging yet vital aspect of feedback communities since the face-value of the information may lead to misleading inferences from users. A common summary statistic used by feedback communities is the average of all ratings. Conventional statistics argue that for mean inference to occur we require a unimodal or symmetrically bimodal distribution. However, upon testing the sampling distribution of ratings, we observe it is rather asymmetrically bimodal, resembling a J-curve shape. Through a series of conventional empirical and sentiment analysis on a sample of reviews from Amazon.com, we hypothesize and examine whether engagement self selection bias via reciprocity and social influence bias via disconfirmation and conformity may lead to such distributional distortions.
Firstly, we identified engagement self-selection bias via reciprocity, whereby reviewers who would otherwise not post a review, driven by reciprocity, they post reviews to reward or punish the seller and thus are more likely to rate on the extreme. We find that frequent reviewers are more likely to rate in the middle than infrequent ones, and that rating on the extreme is driven by high review emotional substance. Moreover, we observe that average word count is higher for middle ratings, implying a higher reviewing "cost" to the reviewer in terms of effort and

time. Secondly, we found evidence of social influence via disconfirmation. We detect that prior beliefs about product quality, shaped upon seeing past reviews before purchase, influence not only our purchasing decisions but they shape posterior beliefs and consequently reviewing behaviour. Intuitively, as customers experiences larger deviations between the expected and actual product quality (disconfirmation) they are more likely to post an extreme rating.

7.1 Recommendations for feedback communities

Engagement self-selection bias can potentially be mitigated by review weighting, accounting for the reviewer's frequency of reviewing. Many platforms already observe frequent reviewers, placing them among the first reviews that customer sees. We, however, propose that a similar approach should be adopted when calculating the average rating. By introducing experience weighting of each rating, the distribution of observed ratings might be altered to resemble a less biased distribution. Another approach is to incentivize more customers to review more often. Thus, customers might build a habit of reviewing the products they purchase, building up a larger community of frequent reviewers. This would create a more randomized sample as customer's with any type of experience, either extreme or mediocre, will engage in feedback communities. A more randomized sample will reduce the sampling biases in the sampling distribution of ratings. As more users post their reviews, the effects of reciprocity on the sampling distributions will be mitigated.

7.2 Limitations and future work

Under the assumption that reviewers can perfectly map experience to review, the validity of research is significantly robust. Nonetheless, research on feedback communities suffers from an inherent omitted variable bias. In real-world studies of reviews, researchers can observe a rating and a review but cannot observe a rater's true experience. Without this information, we can analyse data using expected distributions on large samples or use other variables as proxy accounting for the unobserved effects. Moreover, although through our empirical research we hypothesize behavioural biases, we argue that it would be very insightful to run a set of experiments to capture both reviewer's experience and review. In an effort to examine disconfirmation in a greater detail, feedback communities could perform a difference-in-difference experiment, where the feedback community could

send emails to buyers asking them to review. On one sample they could provide the current rating of the product and on the control they could hide it. Such an experiment could help us understand whether past ratings have a direct effect on reviewing behaviour.
Further, the study of disconfirmation might be influenced by the variable definition and the choice of disconfirmation proxy. It would be wise to test our results by defining an alternative measure of disconfirmation and running the same analysis on an identical dataset. This brings us to the limitations of the available datasets. Expected quality should ideally be calculated before the date of purchase rather than the date of the review; however, the date of purchase is not available to us. Research would also benefit from reviewer demographic characteristics to control for various character reviewer types. Lastly, due to many observations, even simple statistical processes are often too complex and timely to be performed on the full data. It would require a specially designed machine to perform a thorough analysis of the full dataset.
Provided that our empirical analysis was performed on Amazon.com, external validity of our results should be tested on other platforms. This builds on the work of Shen *et al.* (2013), who compare results from Amazon.com and barnesandnoble.com and finds that different platforms may result in different reviewing attitudes. Specifically, review systems and structures differ noticeably between platforms. Therefore, our findings may not apply to reviews from, say, booking.com, which measures satisfaction on a 10 rather than 5 points scale. Lastly, academic controversy between Armstrong (1981) and Liddell and Kruschke (2018) on the validity of mean inference from ordinal data highlights the need of further research in this field.

References

Acemoglu, D., Makhdoumi, A., Malekian, A. and Ozdaglar, A. (2017). *Fast and slow learning from reviews*. Tech. rep., National Bureau of Economic Research.

Adomavicius, G., Bockstedt, J. C., Curley, S. P. and Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, **29** (1), 84–102.

Allison, P. (2008). Convergence failures in logistic regression. *SAS Global Forum 2008*, **360**.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological methods & research*, **28** (2), 186–208.

Anderson, C. J. (2010). Central limit theorem. *The Corsini Encyclopedia of Psychology*, pp. 1–2.

Anderson, J. A. *et al.* (1994). Point biserial correlation. *Stata Technical Bulletin*, **3** (17).

Armstrong, G. D. (1981). Parametric statistics and ordinal data: A pervasive misconception. *Nursing Research*, **30** (1), 60–62.

Babić Rosario, A., Sotgiu, F., De Valck, K. and Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, **53** (3), 297–318.

Basuroy, S., Chatterjee, S. and Ravid, S. A. (2003). How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of marketing*, **67** (4), 103–117.

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, **102** (5), 841–877.

Brynjolfsson, E., Hu, Y. J. and Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. *Sloan Management Review*, **47** (4), 67–71.

Che, Y.-K. and Hörner, J. (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, **133** (2), 871–925.

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

Couture-Beil, A. and Couture-Beil, M. A. (2018). Package 'rjson'. *URL: https://cran. r-project. org/web/packages/rjson/rjson. pdf*.

Cox, N. (2007). Hsmode: Stata module to calculate half-sample modes.

Cox, R. P. . N. J. (2011). "moss: Stata module to find multiple occurrences of substrings". *Statistical Software Components S457261, Boston College Department of Economics, revised 29 Apr 2016.*

Dellarocas, C., Narayan, R. *et al.* (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical science*, **21** (2), 277–285.

Etumnu, C., Foster, K., Widmar, N. O., Lusk, J. and Ortega, D. (2019). Does the distribution of online ratings affect sales? evidence from amazon.

Faber, J. and Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental press journal of orthodontics*, **19** (4), 27–29.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and economic behavior*, **54** (2), 293–315.

Fleder, D. M. and Hosanagar, K. (2007). Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*, pp. 192–199.

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E.*

Greene, W. H. (2000). Econometric analysis 4th edition. *International edition, New Jersey: Prentice Hall*, pp. 201–215.

Grilli, L. and Rampichini, C. (2014). *Ordered Logit Model*, Dordrecht: Springer Netherlands, pp. 4510–4513.

Hecke, T. V. (2012). Power study of anova versus kruskal-wallis test. *Journal of Statistics and Management Systems*, **15** (2-3), 241–247.

Heckman, J. J. (1990). Selection bias and self-selection. In *Econometrics*, Springer, pp. 201–224.

Ho, Y.-C., Wu, J. and Tan, Y. (2017). Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research*, **28** (3), 626–642.

Hu, N., Liu, L. and Zhang, J. J. (2008). Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and management*, **9** (3), 201–214.

Hu, Z. J., N. and Pavlou (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM, 52(10)*, p. 144–147.

Hu Nan, Z. J., Paul A Pavlou (2006). Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, pp. 324–330.

Ifrach, B., Maglaras, C., Scarsini, M. and Zseleva, A. (2019). Bayesian social learning from consumer reviews. *Operations Research*, **67** (5), 1209–1221.

Jann, B. (2019). Estout: Stata module to make regression tables.

Jockers, M. L. (2015). *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

Kahneman, D., Knetsch, J. L. and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, **5** (1), 193–206.

Kaushik, K., Mishra, R., Rana, N. P. and Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on amazon. in. *Journal of retailing and Consumer Services*, **45**, 21–32.

Kawamura, K. (2011). A model of public consultation: why is binary communication so common? *The Economic Journal*, **121** (553), 819–842.

— (2013). Eliciting information from a large population. *Journal of Public Economics*, **103**, 44–54.

Keser, C. (2003). Trust in a networked world: Experimental games for the design of reputation management systems.

Kim, O. and Walker, M. (1984). The free rider problem: Experimental evidence. *Public choice*, **43** (1), 3–24.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, **4**, 83–91.

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, **33** (2), 188–229.

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

Li, H., Xie, K. L. and Zhang, Z. (2020). The effects of consumer experience and disconfirmation on the timing of online review: Field evidence from the restaurant business. *International Journal of Hospitality Management*, **84**, 102344.

Liddell, T. M. and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, **79**, 328–348.

Lin, M., Lucas Jr, H. C. and Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, **24** (4), 906–917.

Long, J. S. and Freese, J. (2001). Fitstat: Stata module to compute fit statistics for single equation regression models.

López, M. and Sicilia, M. (2013). How wom marketing contributes to new product adoption: testing competitive communication strategies. *European Journal of Marketing*.

Luca, M. (2017). Designing online marketplaces: Trust and reputation mechanisms. *Innovation Policy and the Economy*, **17** (1), 77–93.

Marinescu, I., Klein, N., Chamberlain, A. and Smart, M. (2018). *Incentives can reduce bias in online reviews*. Tech. rep., National Bureau of Economic Research.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics*, **3** (4), 303–328.

McKight, P. E. and Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, pp. 1–1.

McKinney, W. *et al.* (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, **14** (9), 1–9.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, **29** (3), 436–465.

Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pp. 195–204.

Ni, L. J., J. and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 188–197.

O'Donovan, J. and Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 167–174.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology*, **62** (4), 480.

Peterson, B. and Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **39** (2), 205–217.

Picard, R., Cox, N. and Ferrer, R. (2017). Rangestat: Stata module to generate statistics using observations within range.

Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pp. 59–72.

Powell, D., Yu, J., DeWolf, M. and Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological science*, **28** (10), 1432–1442.

Rashotte, L. (2007). Social influence. *The Blackwell encyclopedia of sociology*.

Rinker, T. W. (2019). *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York, version 2.7.1.

Shen, W., Hu, Y. and Rees, J. (2013). *Competing for Attention: An Empirical Study of Online Reviewers' Strategic*. Tech. rep., Working paper.

Sherif, M., Taub, D. and Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of experimental psychology*, **55** (2), 150.

Sridhar, S. and Srinivasan, R. (2012). Social influence effects in online product ratings. *Journal of Marketing*, **76** (5), 70–88.

Stevens, S. S. *et al.* (1946). On the theory of scales of measurement.

Team, R. C., Bivand, R., Carey, V. J., DebRoy, S., Eglen, S., Guha, R., Herbrandt, S., Lewin-Koh, N., Myatt, M., Nelson, M. *et al.* (2020). Package 'foreign'.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, **185** (4157), 1124–1131.

Ullah, R., Amblee, N., Kim, W. and Lee, H. (2016). From valence to emotions: Exploring the distribution of emotions in online product reviews. *Decision Support Systems*, **81**, 41–53.

Williamson, O. E. (1979). Transaction-cost economics: the governance of contractual relations. *The journal of Law and Economics*, **22** (2), 233–261.

Wright, R. E. (1995). Logistic regression.

A Appendix

A.1 Programming Languages

To perform our analysis, we primarily used Stata statistical software. Our original data was in JSON data format, and we were able to process them into Stata format using R programming language. We have also utilised R when extracting the price of the product from our meta datasets. Lastly, we used R to perform sentiment and text analysis on our datasets. Python programming language was used to generate the gender of the reviewer from the available reviewer names users provide in their Amazon accounts.

A.1.1 Packages used in Stata

- moss by Cox (2011) to count the number of occurrences of specific words within the text of the reviews.
- csgof to perform chi-square goodness of fit.
- rangestat by Picard *et al.* (2017) to generate statistics within range.
- hsmode by Cox (2007) to generate mode using half-sample estimation.
- estout by Jann (2019) to store and generate regression tables in LaTeX.
- fitstat by Long and Freese (2001) to produce a variety of post-estimation measures-of-fit.
- pbis by Anderson *et al.* (1994) to produce point biserial correlation.

A.1.2 Packages used in R

- ndjson & rjson by Couture-Beil and Couture-Beil (2018) to convert JSON objects to R.
- foreign by Team *et al.* (2020) to convert R data-types to Stata.
- sentimentr by Rinker (2019) to extract sentiment scores.
- syuzhet by Jockers (2015) for emotion and positive/negative words classification.

A.1.3 Packages used in Python

- pandas by McKinney *et al.* (2011) to work and process our datasets.
- json to convert original JSON data-types to delimited CSV files.
- Gender-guesser 0.4.0:This package helps generate the gender of the reviewer given the name of the reviewer as an input. The package uses data from gender program by Jorg Michael to associate names to a gender according to a corresponding likelihood. This likelihood is generated by looking at historical data on names and their respective association to a gender.

A.2 Plots and additional regression output

For a robustness check and to cross-validate our results, we re-perform our primary results using different models and datasets. Figure 11 reproduces Figure 8 on the Alpha dataset. Tables 5 & 6 reproduce Tables 3 and 4 using ordered probit and probit regression models rather than ordered logistic and logistic regression models.

Figure 11: Scatter plot showing the average *extreme* rating of each reviewer against the number of reviews each reviewer has. **(Alpha dataset)**

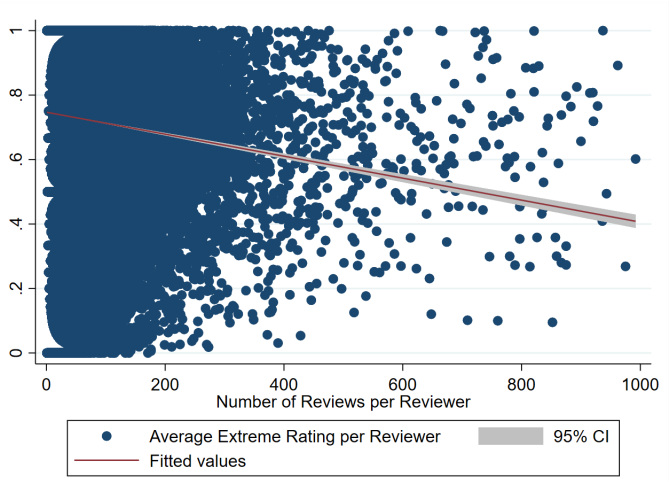


Table 5: Probit regression of *extreme ratings* on *primary variables*, *yearly dummies* & *extensive set of control variables*. **(Gamma Dataset)**

	(1) Model 1	(2) Model 2	(3) Model 3
Absolute Disconfirmation	0.0181	0.0538***	0.0473***
	(0.00975)	(0.00936)	(0.0126)
$Disconfirmation^2$	0.0426***	0.00463	0.00568
	(0.0127)	(0.0123)	(0.0168)
Expected Quality	0.238***	0.225***	0.207***
	(0.00325)	(0.00323)	(0.00339)
Yearly Dummies	-	-	-
	(.)	(.)	(.)
Absolute Sentiment Score			0.270***
			(0.00441)
Observations	614377	614377	250946

Marginal effects; Standard errors in parentheses
(d) for discrete change of dummy variable from 0 to 1
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Ordered probit regression of *ratings* on *primary variables* & *yearly dummies*. (Model 3) **(Gamma Dataset)**

	(1) Rating 1	(2) Rating 2	(3) Rating 3	(4) Rating 4	(5) Rating 5
Absolute Disconfirmation	-0.0102***	-0.00649***	-0.0120***	-0.0195***	0.0481***
	(0.00232)	(0.00149)	(0.00274)	(0.00447)	(0.0110)
$Disconfirmation^2$	-0.0231***	-0.0147***	-0.0271***	-0.0443***	0.109***
	(0.00296)	(0.00189)	(0.00345)	(0.00562)	(0.0139)
Expected Quality	-0.0688***	-0.0440***	-0.0810***	-0.132***	0.326***
	(0.000891)	(0.000590)	(0.000902)	(0.00120)	(0.00254)
Yearly Dummies	-	-	-	-	-
	(.)	(.)	(.)	(.)	(.)
Absolute Sentiment Score	-0.0769***	-0.0492***	-0.0905***	-0.148***	0.364***
	(0.00127)	(0.000897)	(0.00135)	(0.00186)	(0.00460)
Observations	250946	250946	250946	250946	250946

Marginal effects; Standard errors in parentheses
(d) for discrete change of dummy variable from 0 to 1
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B List of variables

Table 7: Key description of variables used in analysis

Variable name	**Variable description**
(Quality) disconfirmation	Actual product quality - expected product quality
Expected quality	Average rating of the product prior user's review
Actual quality	The overall average rating of the product
Asin	Amazon standard identification number (product id)
Number of reviews per reviewer	Number of reviews per reviewer
Number of past reviews	Number of reviews of the product prior user's review
Verified	Binary variable identifying a review, for which reviewer has definitely purchased the product at non discounted price.
Number of reviews per product	Number of reviews per product
Helpfulness votes	Number of helpfulness votes a review received from other users
Price	Price of the product in US$
Female	Binary variable identifying female gender.
Sentiment score	Measure of polarity (emotion) of the review text
Word count of review (summary)	Number of words in the review text or in the summary text
Rating	Takes values: $r \in [1, 2, 3, 4, 5]$
Extreme rating	Binary variable taking value of 1 if $r = 1$ & $r = 5$
Average extreme rating per reviewer	Number of extreme ratings per reviewer divided by the number of all ratings per reviewer
Average rating per reviewer	Mean of all ratings per reviewer
Category	Category that the reviewed product belongs to

C List of figures

Table 8: Description of figures used in our analysis

Figure	Description	Dataset
Figure 1	Words emotion classification of reviews.	Gamma Dataset
Figure 2	Histogram of sentiment score for reviews.	Gamma Dataset
Figure 3	Yearly trend for average rating across all product categories.	Beta Dataset
Figure 4	Sampling distribution of ratings.	Alpha Dataset
Figure 5	Chi-squared probability plot (4 d.f.), for the sampling distribution of ratings.	Alpha Dataset
Figure 6	Normal quantile-quantile probability plot for the sampling distribution of ratings.	Alpha Dataset
Figure 7	Scatter plot showing the average rating of each reviewer against the number of reviews each reviewer has.	Beta Dataset
Figure 8	Scatter plot for average extreme rating of each reviewer against the number of reviews each reviewer has.	Beta Dataset
Figure 9	The average sentiment absolute score, emotion proportion and average review word in each rating.	Gamma Dataset
Figure 10	Sampling distribution of ratings over categories.	Beta Dataset
Figure 11	Scatter plot for average extreme rating of each reviewer against the number of reviews each reviewer has.	Alpha Dataset