Do Bad Businesses Get Good Reviews? Evidence from Online Review Platforms*

Devesh Raval

Federal Trade Commission

draval@ftc.gov

September 21, 2020

Abstract

I examine how review ratings vary with quality across several online platforms, estimating quality tiers for businesses through a finite mixture model using signals from consumer protection authorities and review ratings. Review ratings are higher for Google and Facebook compared to the BBB and Yelp, with larger differences for low quality businesses. Reviews that are likely fake based on multiple review filtering algorithms increase ratings for low quality businesses. Through a linear decomposition, I show that fake reviews can account for about half of the higher average rating for low quality businesses on Google compared to Yelp.

Keywords: online reviews, online platforms, user generated content, consumer protection

*The views expressed in this article are those of the author. They do not necessarily represent those of the Federal Trade Commission or any of its Commissioners. I would like to especially thank Matt Scandale of the Better Business Bureaus for all of his help on this paper. In addition, I would like to thank Carl Bialik and Luther Lowe of Yelp for providing data on Yelp reviews, and Scott Syms for his excellent research assistance. I would also like to thank Chris Adams, Keith Anderson, Mike Atleson, Joe Calandrino, Dan Hosken, Ginger Jin, Yesim Orhun, Jan Pappalardo, Rubens Pessanha, Ted Rosenbaum, Dave Schmidt, Andrew Sweeting, Charles Taragin, Will Violette, and Mo Xiao for their comments on this paper.

1 Introduction

The rise of the Internet has given consumers a megaphone to express their opinions through online reviews. Consumers consult online reviews when making purchasing decisions, while platforms use online reviews to rank products to display to consumers. Online reviews allow businesses to gain a reputation and facilitate trust in the marketplace (Tadelis, 2016).

However, fake reviews may distort the signal of quality in online reviews, and so erode the reputation and trust that they generate. Because online reviews matter so much for consumer decisions (Luca, 2011; Lewis and Zervas, 2020), firms have strong incentives to fake reviews, either to benefit their own business or to hurt rivals.[1] The recent high profile *Sunday Riley* case provides an illustrative example of such behavior; the Federal Trade Commission (FTC) alleged that a company's CEO wrote, and ordered employees to write, 5 star reviews of the company's products using false identities.[2] Indeed, most consumers now believe that they have read fake reviews online (Murphy, 2019).[3]

While online platforms have access to many signals of whether a review is fake, such as online traffic patterns and user activity, the algorithms they use to filter fake reviews are proprietary. Thus, consumers, researchers, and policymakers typically do not know which reviews are fake; it remains unclear how prevalent fake reviews are, or how their prevalence varies across competing platforms.

[1] Luca (2011) finds a 5 to 9% increase in restaurant revenue after a 1 star increase in Yelp rating, and Lewis and Zervas (2020) find a 28% increase in hotel demand after a 1 star increase in rating. For additional evidence, see Chevalier and Mayzlin (2006) on Amazon.com, as well as Resnick et al. (2006) and Cabral and Hortacsu (2010) on eBay.

[2] See https://www.ftc.gov/news-events/press-releases/2019/10/devumi-owner-ceo-settle-ftc-charges-they-sold-fake-indicators.

[3] Murphy (2019) find that 82% of consumers surveyed in 2019 report reading a fake review, and 24% were asked by a business to write a review in exchange for cash, freebies, or discounts.

In addition, it is theoretically ambiguous whether high quality or low quality firms are more likely to fake reviews, as well as whether greater prevalence of fake reviews raises or lowers welfare (Dellarocas, 2006). Thus, it is also unclear whether fake reviews distort the measure of quality that consumer reviews provide.

In this article, I directly estimate a firm's quality using several measures of the likelihood of potential consumer protection problems, together with online review ratings. I then examine how consumer generated reviews vary across platforms. I find systematic differences in scores across platforms, with the largest differences for the lowest quality firms using my measure of quality. I then show that fake reviews likely account for some, but not all, of these differences across platforms.

I examine five platforms – the Better Business Bureau (BBB), Yelp, Google, Facebook, and HomeAdvisor – by matching a sample of over one hundred thousand businesses to review listings. While Google and Facebook have become dominant platforms for online reviews, they have so far not featured in the economics and marketing literature on online reviews.

Despite the fact that all of these platforms measure online ratings on a 5 point scale, the distribution of review ratings is very different across platforms. BBB average ratings are bimodal with most ratings either very low or very high, while Yelp ratings are much more uniform across the rating distribution. Facebook, Google, and HomeAdvisor have heavily skewed distributions with most businesses having very high ratings. In addition, businesses with signals of poor quality, such as a large number of complaints or a F grade from the BBB, have lower average ratings on the BBB and Yelp compared to Google, Facebook, and HomeAdvisor.

I use these signals of quality to structurally estimate quality tiers through a non-parametric finite mixture model. Non-parametric identification requires at least three signals that are independent conditional on the quality type; I have ten signals. The first set of signals is based upon consumer complaints to consumer protection organizations, and includes the BBB letter grade of the business, data on current and past complaints, and their successful resolution of these complaints. The second set of signals are the review ratings of the five platforms.

I estimate three quality tiers; businesses in the high quality tier have almost no complaints and almost all receive A+ grades from the BBB. The low quality tier includes about 10% of businesses in the sample; these businesses receive a large number of complaints and are more likely to receive a F grade from the BBB. In addition, low quality businesses are much more likely to be designated as high risk for fraud by the BBB, a measure that is not used in model estimation. Thus, the estimated quality tiers reflect the likelihood of experiencing consumer protection issues.

Both low and high quality businesses have higher ratings on Google, Facebook, and HomeAdvisor compared to the BBB and Yelp. However, the difference between platforms is much larger for low quality businesses. On average, Google ratings are about a half star higher than Yelp for high quality businesses, but about a star higher for low quality businesses. In contrast, relative rankings, which might affect platform search results, are fairly consistent across platforms; for all platforms, low quality businesses almost always have a lower rating than high quality businesses.

Fake reviews can likely explain some of the differences in ratings between platforms for low quality businesses. I examine fake reviews through proxies for whether a review is fake. For Yelp, I know whether the review is "hidden" from view on Yelp because it is flagged by

Yelp's algorithms as likely fake. For the BBB, I have the score from a proprietary filtering algorithm predicting the probability that a review is fake. For both the BBB and Yelp, the share of likely fake reviews is similar for high and low quality businesses. However, ratings of reviews that are likely to be fake are substantially higher than ratings from published reviews for low quality businesses. That is, other things equal, likely fake reviews of low quality businesses are higher than those of likely true reviews. By contrast, the difference for high quality businesses is much smaller for Yelp, and negligible for the BBB.

On the other hand, platforms differ in ratings for reasons beyond fake reviews, such as the type of reviews posted. Unlike Yelp, Google allows "no-text" reviews. Examining two large ancillary datasets of reviews, I find that 45% of Google reviews are 100 characters or less, many of which are no-text reviews, while only 4% of Yelp reviews are 100 characters or less. Because longer reviews tend to be more negative, differences in review length can account for part of the lower average review rating of Yelp compared to Google.

I then develop a linear decomposition to assess how much of the difference in ratings between platforms is due to fake reviews, using a "difference in difference" approach comparing high and low quality businesses between Google and Yelp. If fake reviews do not differentially affect ratings of high quality businesses across platforms, about half of the difference in the average rating of low quality businesses between Google and Yelp is due to fake reviews. The share of reviews of low quality businesses on Google that are fake is not identified, but a lower bound is that 30% of low quality business reviews on Google are fake.

Given this evidence that fake reviews steer consumers to low quality businesses to their detriment, policymakers' recent efforts to police fake reviews likely improve welfare. The Competition and Markets Authority (CMA) of the UK has launched an investigation into fake reviews on several online platforms, while the FTC has now brought several cases alleging review manipulation by a business on an online platform. This research also answers FTC Commissioners Rohit Chopra and Rebecca Slaughter's call for the FTC to "comprehensively study the problem of fake reviews, given their distortionary effect on our markets".[4]

In addition, when regulating platforms, regulators may have to consider how changes in platform conduct with respect to online reviews would affect user welfare. For example, the FTC previously investigated allegations that Google's changes to its search results to favor its own reviews over competitors such as Yelp were anticompetitive.[5]

A recent literature in economics and marketing has examined the issue of fake reviews. Mayzlin et al. (2014) identify fake reviews through differences in platforms, comparing verified reviews on Expedia to unverified reviews on TripAdvisor. This article builds on their work by examining online reviews across several platforms. Several additional articles use evidence of fake reviews for a single platform, either using filtered reviews on Yelp (Luca and Zervas, 2016), reviews with no record of purchase for a private label retailer (Anderson and Simester, 2014), or records of purchased reviews on Amazon from Facebook groups of fake review buyers (He et al., 2020).

An extensive literature in computer science has also examined fake reviews (Kumar and

[4]See https://www.ftc.gov/system/files/documents/public_statements/1550127/192_3008_final_rc_statement_on_sunday_riley.pdf for the Commissioners' statement. Additional FTC cases on fake reviews include Cure Encapsulations, Urthbox, Mikey & Momo (aka Aromaflage), Universal City Nissan (aka Sage Auto), Son Le and Bao Le (aka Trampoline Safety of America), Amerifreight, and LendEDU. For the CMA investigation, see https://www.gov.uk/government/news/cma-investigates-misleading-online-reviews. The Australian Competition and Consumer Commission (ACCC) also recently won a case against a website accused of creating fake reviews; see https://www.accc.gov.au/media-release/service-seeking-to-pay-penalty-for-misleading-online-customer-reviews.

[5]The FTC, after closing the investigation, stated that "Google adopted the design changes that the Commission investigated to improve the quality of its search results, and that any negative impact on actual or potential competitors was incidental to that purpose". See https://www.ftc.gov/sites/default/files/documents/public_statements/statement-commission-regarding-googles-search-practices/130103brillgooglesearchstmt.pdf.

Shah, 2018), focusing on identifying ways to detect fake reviews (Plotkina et al., 2020; Rayana and Akoglu, 2015; Shehnepoor et al., 2017; Wu et al., 2010; Ye et al., 2016), as well as evaluating the effectiveness of fake review attacks (Lappas et al., 2016).

This article is also related to work on signals of business quality and their effects on markets. Jin and Leslie (2003) show that releasing restaurant grades improves restaurant hygiene quality. Jin and Kato (2006) examine trading cards on eBay, and find that neither seller ratings or seller claims provide a complete guide to product quality, with some sellers committing fraud. Tadelis and Zettelmeyer (2015) show that disclosing information on quality increases seller revenue, even when quality is low, as information disclosure improves matching. De Langhe et al. (2016) compare a traditional measure of quality – Consumer Reports reviews – to Amazon product reviews, and find little correlation between the two.

Section 2 details how I construct a dataset of review ratings across several platforms. Section 3 provides evidence of the correlation between review ratings and multiple signals of business quality, while Section 4 uses these signals to estimate a finite mixture model of business quality. Section 5 examines fake reviews as an explanation for my findings, and Section 6 concludes.

2 Data

2.1 Sample Construction

The BBB provided me with data from their database of businesses as of February 28, 2020, which included more than 4.8 million unique businesses. My goal was to examine businesses with any recent activity on the BBB's platform. Thus, I define the sampling frame to include only US located businesses with a BBB letter grade, and with at least one review or one complaint within three years.

In addition, I removed businesses that might match many different listings on a review platform. For example, the BBB's listing of Citibank would be its corporate headquarters, while Google or Yelp would have review listings at the bank branch level for thousands of branches. In order to exclude such headquarters' listings, the sample was restricted to businesses with 1,000 employees or less, and with fewer than 6 listed locations. The sampling frame then has 628,478 businesses.[6]

Because of financial constraints for the Google APIs described in the next section, I could not match all of these businesses to listings on review platforms. I thus examine a random sample of businesses developed through a stratified sampling design. My approach in the stratified sampling design was to oversample businesses with either significant activity on the website or that likely have consumer protection problems.

Table I details the sampling design, including the total sample size, universe size, and the probability of selection for each group.[7] I first proxy for significant activity on the BBB's website as businesses with at least 10 or more reviews or at least 10 or more complaints; I sample all such businesses.

In addition, I oversample businesses with two indicators of potential consumer protection problems. The BBB's line of business designation has several categories that the BBB considers high risk, such as ponzi schemes, prize promotions, and advance fee brokers. I

[6] I also excluded a small number of businesses with a "#" in their name, as this interferes with the API calls described in the next subsection.

[7] The sampling weight is the inverse of the probability of selection.

Table I Sampling Design

Sampling Group	In Sample	In Universe	Selection Probability
BBB Complaints $\geq$ 10 or Reviews $\geq$ 10	32,641	32,641	100%
Business in High Risk Category	1,723	1,723	100%
BBB Letter Grade C+ or Below	50,000	132,175	37.8%
BBB Letter Grade B- or Above	50,000	461,939	10.8%
Total	134,364	628,478	

Note: The groups of BBB Letter Grade C+ and Below and BBB Letter Grade of B- and Above are based on all businesses with less than 10 BBB reviews and less than 10 BBB complaints, and not designated in a high risk category. The number of BBB reviews and complaints are based on a three year window.

sampled all businesses in the sampling frame designated by the BBB as high risk for fraud. Of the remaining businesses, I stratify sample based on the BBB letter grade of business quality, which ranges from A+ to F. I divide businesses into those with a high grade (B- or better) or low grade (C+ or worse), and randomly sampled 50,000 businesses with a B- grade or better and 50,000 businesses with a C+ grade or worse. Because the sampling frame has many more high grade businesses than low grade businesses, 37.8% of high grade businesses are sampled, compared to 10.8% of low grade business.

2.2 Review Platform Data

The BBB provided me with data from their reviews. I then matched the sample businesses to review ratings from other platforms through two Google APIs. Each API provided the average rating for a business and the number of customer reviews for that business.

The first API, the Google Search API, provides Google Custom Search results for search queries. For each business, I used a search string of the business name, city, state, and zip code.[8] The API provides details on the review rating and number of reviews for business listings on review platforms in the top 10 Google search results. I include ratings from the

[8]Because of usage limits, I had to space out API queries over a fortnight.

three platforms with the largest number of businesses with review ratings: Yelp, Facebook, and HomeAdvisor. Yelp is a major general review platform, Facebook the dominant social media platform, and HomeAdvisor a large specialty platform in home improvement and repair professionals.[9]

The Google Search API does not provide data on business listings on Google's own platform. Thus, I also used the Google Places API in order to obtain Google review ratings using the same search string of the business name, city, state, and zip code.

Next, I cleaned the data by comparing the business name, street address, and zip code provided in the API results to those in the BBB data; Appendix A provides details on this matching process. I required the business name, street address, and zip code to all match within a specified tolerance in order to be included in the final dataset. A research assistant then compared matched listings with the platform websites for a random sample of listings; 99.5% of Google listings, 99.5% of Yelp listings, 96.9% of Facebook listings, and 99.0% of HomeAdvisor listings are coded as correctly matched.[10] Finally, I matched data on Yelp listings to data on individual reviews (both published and hidden) provided to me by Yelp using the business's website link.

[9] While Angie's List also had a large number of review listings, the Google Search API did not provide review ratings; HomeAdvisor's parent IAC recently purchased Angie's List and combined the two companies into ANGI Homeservices Inc. However, both brands continue to have separate review listings.

[10] I demonstrate in Appendix A that the accuracy of the match declines if I make the matching criteria less stringent. Facebook listings in particular tend to be harder to match than Google, Yelp, or HomeAdvisor listings, because the Google Search API does not always record the full address, or record the address in a consistent format.

Table II Origin of Variables Used in the Analysis

Variable	Origin
Review Ratings	
BBB Reviews and Ratings	Provided by BBB
Yelp Reviews and Ratings	Google Search API Matched to Yelp Internal Data
Google Average Ratings	Google Places API
Facebook Average Ratings	Google Search API
HomeAdvisor Average Ratings	Google Search API
Consumer Protection Signals	
BBB Letter Grade	Provided by BBB
BBB Complaints	Provided by BBB
Non-BBB Consumer Sentinel Complaints	Matched From Consumer Sentinel Network

2.3 Final Dataset

Finally, I match the sample to several signals from consumer protection authorities. First, the BBB provided data on consumer complaints from 2017 to 2020 and the BBB letter grade of the business, which I used to construct the sample. Later on, I also matched the sample to data from the BBB on whether complaints were resolved, as well as complaint data from 2010 to 2016.[11] In addition, I match this data to data on non-BBB complaints from January 2015 to April 2020 to the Consumer Sentinel Network, a large database of complaints to the FTC, other federal agencies such as the CFPB, state agencies, and other organizations.[12] Table II provides information on the main variables used, as well as their origin.

In the resulting dataset, the median number of BBB complaints is one and the mean number of BBB complaints is 2.7; 81% of businesses have at least one BBB complaint. Examining the BBB letter grade of the business, 54% of businesses have an A+ grade and

[11] When using these variables, I have to exclude 324 businesses for which I cannot match these complaint measures to the original dataset.

[12] See Raval (2019) for more details on the Consumer Sentinel Network. I fuzzy matched complaints from Consumer Sentinel to each business based on the business name and zip code; the name was matched using a Jaro-Winkler distance with $p = 0$ and a threshold of 0.125 on the distance metric, and exact matching on the zip code.

7% have an F grade.[13] The share of businesses with a matched review listing is highest for the BBB and Google: 38% of businesses in the full sample have at least one BBB review listing, compared to 47% for Google, 19% for Yelp, 9% for Facebook, and 2% for HomeAdvisor.

3 Ratings and Signals of Quality

In this section, I first examine the distribution of ratings across platforms. I then examine how ratings correlate with two signals of quality: the BBB letter grade of the business and the number of complaints received.

3.1 Distribution of Ratings: the "Lake Wobegon" Effect

The distribution of average business ratings is quite different across platforms; Figure 1 displays this distribution. The BBB ratings are bimodal, with most businesses having either a rating above 4 stars or below 2 stars. Yelp ratings look much more uniform across the rating distribution. Finally, for Google, Facebook, and HomeAdvisor, most businesses have average ratings above 4 stars. Like all the children of Garrison Keillor's fictional Lake Wobegon, almost all businesses on Google, Facebook, and HomeAdvisor are above average.

Table III displays the mean star rating by platform in the first column, the share of ratings above 4 stars in the second column, and the share of ratings below 2 stars in the third column. The mean BBB rating is 3, the mean Yelp rating 3.4, the mean Google rating 4, the mean Facebook rating 4.4, and the mean HomeAdvisor rating 4.7. On average,

[13] All estimates described in the paragraph weight using the sampling weights. In the unweighted data for the entire sample, the share of F graded businesses is higher and the share of A+ graded businesses in lower, as should be expected given the sampling design – 38% of businesses have a A+ grade and 17% have a F grade.

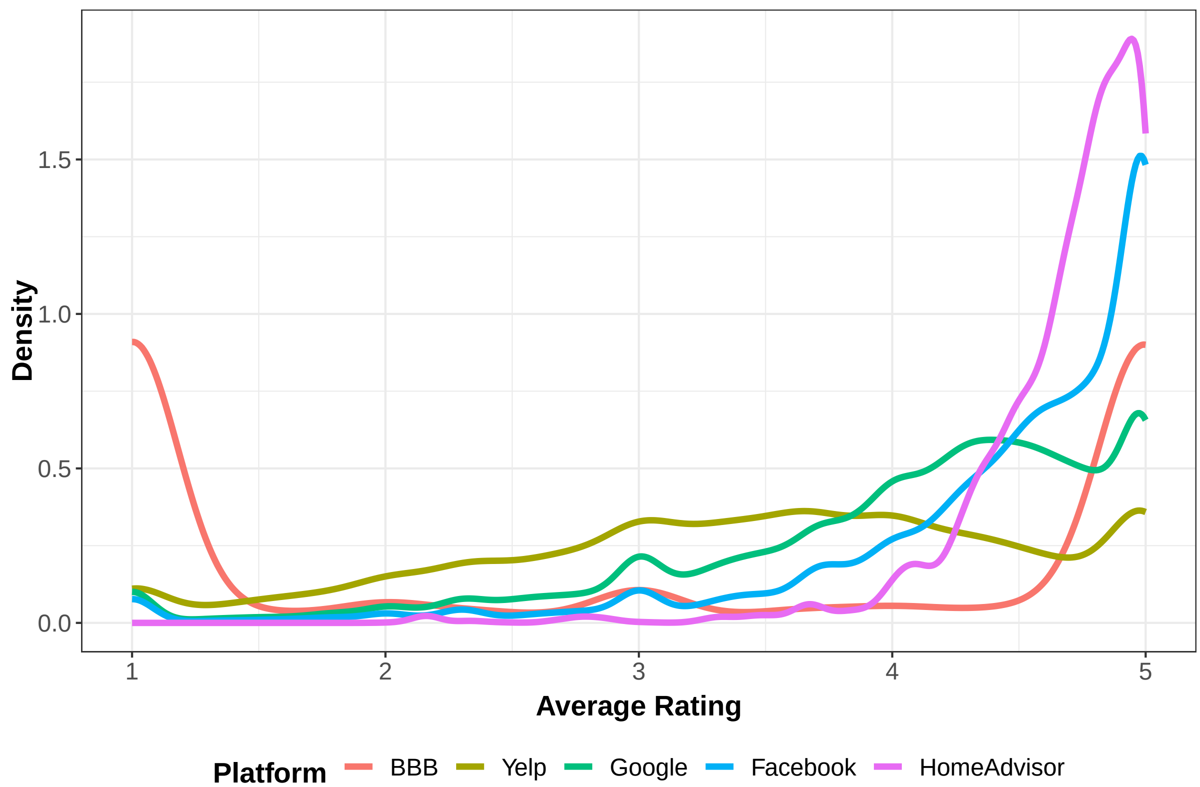


Figure 1 Distribution of Average Business Ratings Across Platforms

Note: All observations weighted using the sampling weights.

32% of Yelp businesses have an average above 4 stars, compared to 44% for the BBB, 59% for Google, 79% for Facebook, and 96% for HomeAdvisor. In contrast, while 10% of Yelp businesses and 43% of BBB businesses have an average below 2 stars, only 4% of Google businesses and 2% of Facebook businesses have an average below 2 stars. No HomeAdvisor businesses have an average below 2 stars.[14]

These differences are not primarily driven by the composition of businesses with review ratings across different platforms. To show this, I control for business fixed effects, which

[14]One reason why HomeAdvisor might have few poorly rated businesses is that businesses on its platform have to pass criminal background and licensing checks. See https://www.homeadvisor.com/screening/. I thus control for individual business fixed effects in many of my specifications.

Table III Star Ratings by Platform

	Mean (1)	Share > 4 (2)	Share < 2 (3)
BBB	3.01	0.44	0.43
	(0.01)	(0.00)	(0.00)
Yelp	3.44	0.32	0.10
	(0.01)	(0.00)	(0.00)
Google	4.01	0.59	0.04
	(0.00)	(0.00)	(0.00)
Facebook	4.39	0.79	0.02
	(0.01)	(0.00)	(0.00)
HomeAdvisor	4.67	0.96	0.00
	(0.01)	(0.00)	(.)
Observations	159257	159257	159257

Note: Estimates clustered at the individual business level and include all businesses in the sample weighted using the sampling weights.

control for *any* differences across businesses, through the following specification:

$$Y_{ip} = \gamma_p + \delta_i + \epsilon_{ip}, \tag{1}$$

where Y_{ip} is either the average rating for business i on platform p, an indicator of whether the rating is above 4 stars, or an indicator of whether the rating is below 2 stars. I include business fixed effects through δ_i, and platform fixed effects, measured relative to the omitted category of the BBB's average review ratings, through γ_p. The platform fixed effects are the object of interest.

Table IV displays these results; Google, Facebook, and HomeAdvisor continue to have substantially higher ratings than either the BBB or Yelp. After controlling for business fixed effects, Yelp's rating is 0.5 stars higher than the BBB's rating on average, Google's rating is 1.2 stars higher on average, Facebook's rating is 1.3 stars higher, and HomeAdvisor's rating is 0.9 stars higher. Google, Facebook, and HomeAdvisor have more ratings greater than 4 stars compared to the BBB, and less ratings lower than 2 stars compared to the BBB. Yelp has less 4 star ratings than the BBB and less 2 star ratings relative to the BBB, consistent with the more uniform distribution across ratings as seen in Figure 1.

Table IV Differences in Star Rating by Platform from BBB Ratings

	Mean (1)	Share > 4 (2)	Share < 2 (3)
Yelp	0.46	-0.11	-0.32
	(0.01)	(0.01)	(0.00)
Google	1.15	0.22	-0.40
	(0.01)	(0.00)	(0.00)
Facebook	1.27	0.32	-0.37
	(0.02)	(0.01)	(0.00)
HomeAdvisor	0.94	0.30	-0.27
	(0.03)	(0.01)	(0.01)
Observations	111323	111323	111323

Note: Estimates clustered at the individual business level and include all businesses in the sample weighted using the sampling weights.

3.2 BBB Letter Grade

Next, I examine how review ratings vary by the BBB letter grade for the business. The BBB assigns grades from A+ to F with plus and minus grades for all letter grades except F; these grades do not depend upon review ratings.[15] For purposes of analysis, these are aggregated into 6 groups: A+, A or A-, any B grade, any C grade, any D grade, or F. Figure 2a depicts the average rating by BBB letter grade.

I find a decline in the average rating with worse BBB letter grades for all five platforms; however, this decline is much larger for the BBB than the other platforms. The average A+ business has a 3.3 star rating for the BBB, compared to a 1.7 star rating for a F business, a

[15] The BBB develops its letter grade based on seventeen factors, many of which depend upon the complaints it receives. Unlike complaints, reviews do not enter into the letter grade calculation. See https://www.bbb.org/overview-of-bbb-ratings and https://www.bbb.org/canton/get-consumer-help/rating-faq/.

decline of 1.58 stars. Ratings decline by about a star between an A or A- graded business and a business with any B grade. The decline in rating from an A+ business to F business is, on average, 0.89 stars for Yelp, 0.76 for Google, 0.56 for Facebook, and 0.27 for HomeAdvisor. On average, a F graded business on Google, Facebook, and HomeAdvisor has a higher rating than a A+ graded business on the BBB!

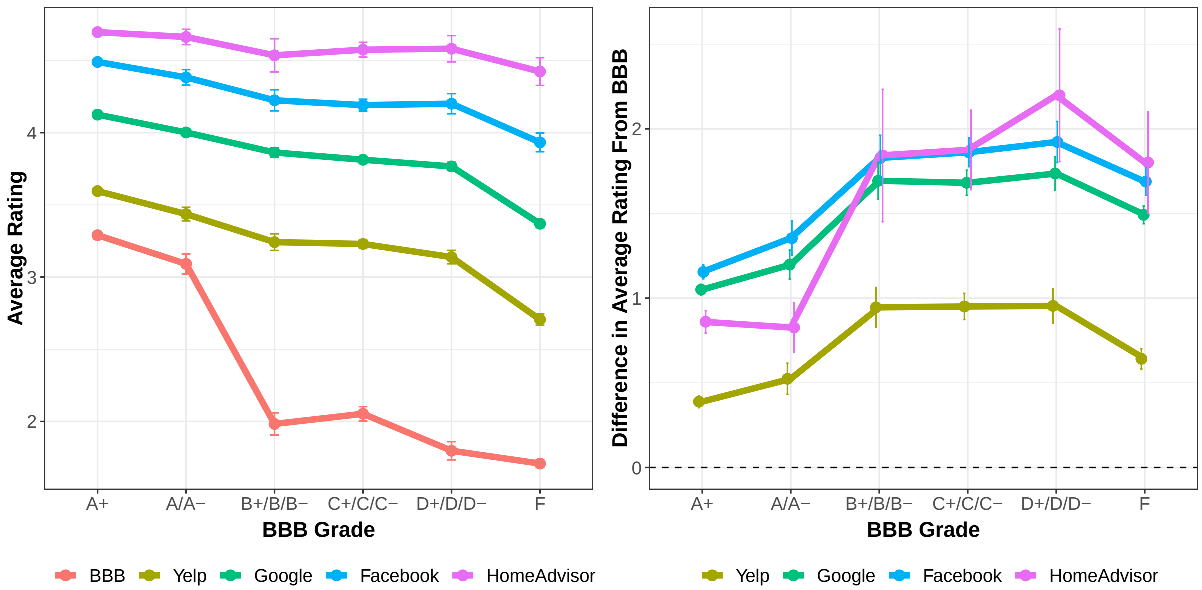
In order to control for differences in the composition of businesses with review ratings across different platforms, I control for business fixed effects by estimating the following specification:

$$Y_{ip} = \beta_{g(i)p} + \delta_i + \epsilon_{ip}, \tag{2}$$

where Y_{ip} is the mean rating for business i on platform p, δ_i are business fixed effects, and $\beta_{g(i)p}$ are indicators for platform p with BBB letter grade $g(i)$, measured relative to the omitted category of the BBB's ratings.

Figure 2b depicts estimates of these specifications. Google, Facebook, and HomeAdvisor have much higher ratings for low letter grade businesses than the BBB, so the gap between these platforms and the BBB rating increases when the letter grade declines. Yelp ratings tend to be closer to the BBB's rating, although the gap between Yelp and BBB ratings also rises as the letter grade declines.

An A+ business has a 0.39 star higher rating on Yelp than the BBB, a 1.05 star higher rating on Google, a 1.15 higher star rating on Facebook, and a 0.86 higher star rating on HomeAdvisor. For F graded businesses, Yelp ratings are 0.64 higher than BBB ratings, Google ratings 1.49 stars higher, Facebook ratings 1.69 stars higher, and HomeAdvisor ratings 1.8 stars higher. The gap between BBB ratings and other platform ratings is thus higher



(a) Average Rating

(b) Rating Relative to BBB

Figure 2 Rating by Platform and BBB Grade

Note: Estimates clustered at the individual business level and include all businesses in the sample weighted using the sampling weights.

for lower letter grade businesses; it grows by 0.25 stars for Yelp, 0.44 stars for Google, 0.53 stars for Facebook, and 0.94 stars for HomeAdvisor.

3.3 BBB Complaints

Finally, I examine how review ratings vary by the number of BBB complaints for the business received in the past 3 years. For purposes of analysis, I group the number of complaints into 6 groups: 0, 1, 2-4, 5-9, 10-24, or 25 or greater complaints. I depict the average rating by the number of complaints in Figure 3a.

The average rating declines with more complaints for all five platforms; however, this decline is much larger for the BBB and Yelp than the other platforms. The average business with zero complaints has a 3.6 star rating for the BBB and 4 star rating for Yelp. A business with 25 or more complaints has, on average, a 1.8 star rating on the BBB and a 2.1 star rating on Yelp, a decline of 1.8 stars for the BBB and 1.9 stars on Yelp when going from 0 complaints to 25 or more complaints. The decline in rating going from 0 complaints to 25 or more complaints is, on average, 1.2 for Google, 1.0 for Facebook, and 0.5 for HomeAdvisor, significantly lower than for the BBB or Yelp.

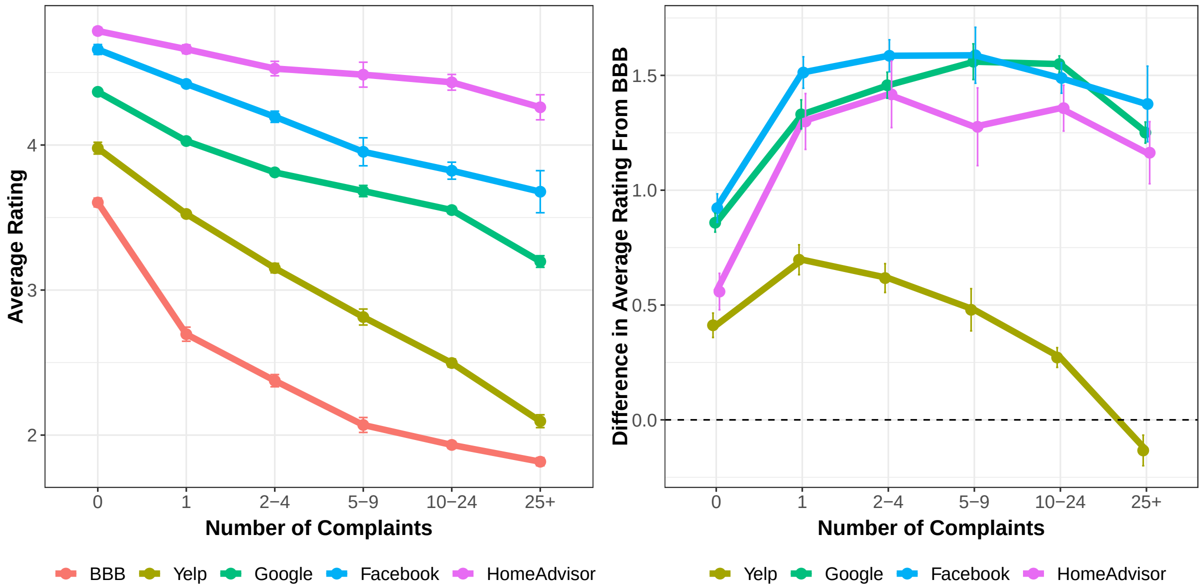
I control for business fixed effects by estimating the following specification:

$$Y_{ip} = \alpha_{c(i)p} + \delta_i + \epsilon_{ip}, \tag{3}$$

where Y_{ip} is the mean rating for business i on platform p, δ_i are business fixed effects, and $\alpha_{c(i)p}$ are indicators for platform p with complaint category $c(i)$, measured relative to the omitted category of the BBB's ratings.

Figure 3b depicts estimates of these specifications. Google, Facebook, and HomeAdvisor have much higher ratings for businesses with many complaints than the BBB, so the gap between these platforms and the BBB rating increases when the number of complaints rises. Yelp ratings tend to be closer to the BBB's rating, and the gap between the BBB rating and Yelp rating declines with more complaints.

A business with zero complaints has a 0.41 star higher rating on Yelp than the BBB, a 0.86 star higher rating on Google, a 0.92 higher star rating on Facebook, and a 0.56 higher star rating on HomeAdvisor. For businesses with 25 or more complaints, Yelp ratings are 0.1 stars lower than BBB ratings. In contrast to Yelp, the gap between the other platforms and the BBB rises; for businesses with 25 complaints or more, Google ratings are 1.25 stars higher than the BBB, Facebook ratings 1.37 stars higher, and HomeAdvisor ratings 1.16



(a) Average Rating

(b) Rating Relative to BBB

Figure 3 Rating by Platform and Number of Complaints

Note: Estimates clustered at the individual business level and include all businesses in the sample weighted using the sampling weights.

stars higher.

4 Estimating Quality

In the previous section, I examined how average review ratings for several platforms varied with two signals of quality: the BBB letter grade of the business and the number of complaints received by the BBB over a three year period. In this section, I structurally estimate quality tiers through finite mixture modeling using several different signals.

4.1 Finite Mixture Models

A finite mixture model assumes that the businesses in the dataset are comprised of a set of unobserved latent classes or types. While the observed data do not identify which businesses are of what type, a finite mixture model helps translate signals into information about the likelihood that the business belongs to a given type. I interpret these groups as quality tiers using how the distribution of signals varies across groups. Because the signals I use are based, in part, on complaints to consumer protection organizations, this measure of quality reflects the likelihood of experiencing consumer protection issues.

A finite mixture model is appropriate for this question for two reasons. First, it fits how review platforms operate quite well. Google and Yelp ask consumers to assign a business a score from one to five (five tiers), while the BBB grade has thirteen tiers in the letter grades from A to F with plus and minus gradations. Second, a major purpose of the BBB letter grade is to inform consumers about businesses with likely consumer protection problems; such businesses might naturally comprise the lowest quality tier. Consumer protection authorities have previously used finite mixture models to identify businesses that were likely to have committed fraud (Balan et al., 2015).

Formally, under the finite mixture model, the likelihood of the data for business i is:

$$L(x_{i1}, x_{i2}, ..., x_{iK}) = \sum_{j=1}^{J} \lambda_j \prod_{k=1}^{K} f_{jk}(x_{ik}), \tag{4}$$

where there are J types in the population with type j having proportion λ_j. The observed data has K quality signals, where x_{ik} is signal k for business i. For each type j, the distributon of signal k is f_{jk}.

For the application in this paper, it is important that the distribution of signal k for type j, f_{jk}, is allowed to be non-parametric. Early work on mixture models had assumed normal signals. However, as Figure 1 demonstrates, the distribution of review ratings is not normal for any of the platforms, and varies considerably across platforms. Similarly, the BBB letter grade is a signal with 13 values, with a mode at the highest grade of A+, and the distribution of complaints has a long tail of businesses with many complaints.

The finite mixture model is non-parametrically identified if there are at least three signals that are independent of each other conditional on the unobserved type (Allman et al., 2009).[16] This identification is up to "relabeling", as the order of the components is not identified. In practice, I use the distribution of signals across types to label these types as quality tiers.

4.2 Quality Estimates

To estimate the non-parametric mixture model, I use the approach of Levine et al. (2011) as implemented in Benaglia et al. (2009b).[17] The Levine et al. (2011) algorithm treats the unobserved types as "missing data", and adopts a majoritization-minorization (MM), or EM-like, iterative approach to estimation. In the majorization step, one estimates the posterior probability that each business is in each type based on the values of the signals, conditional on estimates of the signal distributions f and the type shares λ. In the minorization steps, conditional on the probabilities, one estimates the type shares λ by averaging the posteriors, and the signal distributions f through kernel density estimation. I provide the algorithm steps in Appendix B.1. Levine et al. (2011) prove descent properties for the algorithm, which iterates until convergence. To identify types, I assign each business the type with the highest

[16] Early work by Hall and Zhou (2003) and Hall et al. (2005) had proved that at least three signals were required with two mixture components; Allman et al. (2009)'s proof of the more general case builds on Kruskal (1977). For recent additional work in economics on the identification of mixture models, see Adams (2016) and Kasahara and Shimotsu (2014).

[17] Benaglia et al. (2009b) also includes alternative algorithms from Benaglia et al. (2009a) and Chauveau and Hoang (2016). Levine et al. (2011) show that their algorithm performs similarly to Benaglia et al. (2009a), but it is orders of magnitude faster for my application given the size of my dataset. Hall et al. (2005) and Bonhomme et al. (2016) propose alternative estimation approaches.

posterior probability.

I then estimate a mixture model with ten signals and three types. The first set of signals that I use are explicitly consumer protection related; four come from the BBB and one from the Consumer Sentinel Network. First, I include the two signals examined in Section 3: the BBB letter grade and the number of complaints to the BBB in the last three years. Second, I include the share of such complaints coded by the BBB as not having been resolved, of all unresolved and resolved complaints in the previous three years. Third, I include the number of BBB complaints from 2010 to 2016, a seven year period prior to the complaint measure examined in Section 3. Finally, I include the number of non-BBB complaints from January 2015 to April 2020 to the Consumer Sentinel Network. I include the BBB letter grade as a numeric value from 1 to 13, and all complaint measures as the log of the number of complaints plus one.

The second set of signals are the review ratings from the BBB, Yelp, Google, Facebook, and HomeAdvisor. By also including review ratings as signals, I allow the BBB to be "wrong". For example, a business that has poor reviews on all of the platforms could be placed in a low quality tier, even if the BBB assigns it an A+ letter grade. For review ratings, many businesses will have no rating because they do not have reviews on the platform; I give the signal a value of 10 to indicate such missing values.

In Table V, I provide statistics on the characteristics of each tier; I label these as "high", "medium", and "low" quality tiers. Only about 10% of the businesses are in the low quality tier, compared to 27% in the high tier and 63% in the medium tier.

Almost all the medium and low quality businesses have BBB complaints, compared to only 28% of high quality businesses. However, low quality businesses have many more complaints than medium quality businesses – a median of 6 compared to 1, and a mean of 17.3 compared to 1.4. A substantial number of complaints for both medium and low quality businesses are unresolved as well.

For the BBB letter grade, a large fraction of the low quality businesses are F graded (33%), while few of the high quality (0.5%) or medium quality (6.3%) have F grades. Similarly, almost 92% of high quality business have a BBB grade of A+. However, surprisingly, 40% of medium quality businesses and 34% of low quality businesses have an A+ grade. Low quality businesses with an A+ grade tend to have many complaints but to have successfully resolved almost all of these complaints.

I evaluate the model's performance through a risk measure not included as a signal, the share of businesses deemed as high risk by the BBB (such as pyramid schemes or work at home companies).[18] The low quality tier has the largest share of these businesses, at 1.2%, followed by the medium tier at 0.3% and the high quality tier at 0.1%. Thus, the share of high risk businesses increases when quality falls, as one would expect, even though the model did not explicitly use this designation as a signal.

Figure 4a examines how the average rating of each platform varies by quality tier. For the BBB, the average rating decreases substantially going from the high quality tier to the medium quality tier; the average high quality business has a rating of 3.5 stars, while the average medium quality business has a rating of 1.9 stars and the average low quality business has a rating of 1.8 stars. For Yelp, the decline is much more substantial from medium quality

[18]The full list of categories are: Advance Fee Brokers, Advance Fee Job Listing and Advisory Services, Advance Fee Residential Loan Modification (CA), Chain Letter, Credit Repair Advanced Fee, Deceptive Telemarketing Office Supply Sales, Foreign Lottery, Foreign Online Pharmacy, High Risk Behavior/Practices, High Risk Free Trial Offers, Non-Compliant Debt Relief Services, Online Casino, Paving, Painting, Home Improvement - Itinerant Workers, Ponzi Scheme, Prize Promotions, Pyramid Companies, Reloaders, Sweepstakes, and Work-At-Home Companies.

Table V Summary Statistics by Quality Tier

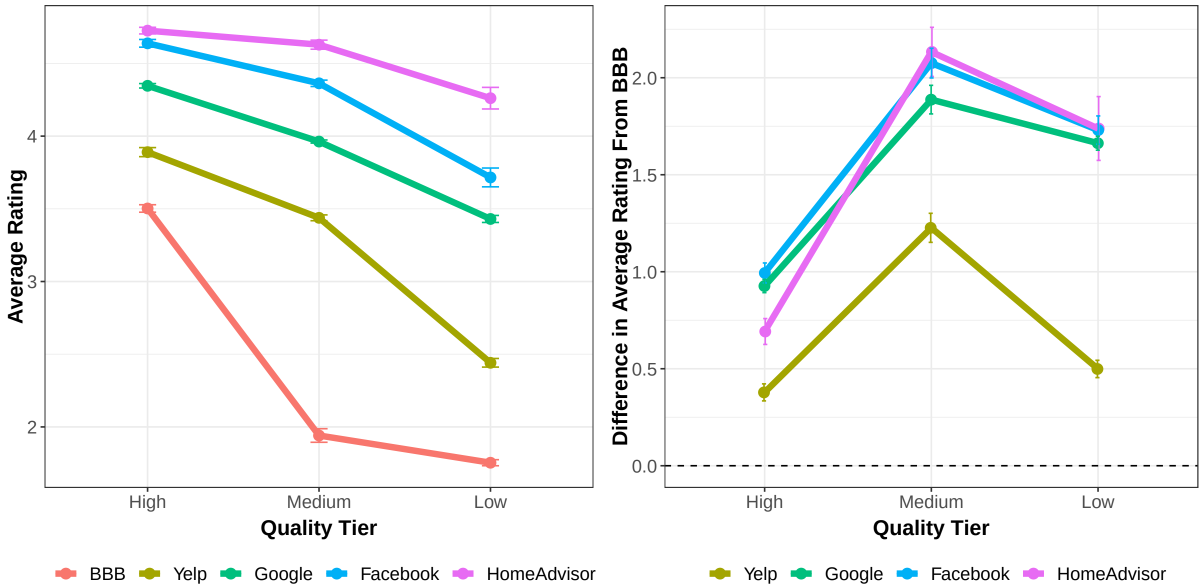
	High	Medium	Low
Type Share	26.9%	63.4%	9.7%
Median Number of Complaints	0.0	1.0	6.0
Mean Number of Complaints	0.5	1.4	17.3
Share High Risk	0.1%	0.3%	1.2%
Share With Complaints	27.8%	100.0%	99.5%
Share with A+ Grade	91.6%	40.2%	34.3%
Share with F Grade	0.5%	6.3%	32.7%
Share of Complaints Unresolved	0.2%	47.7%	32.4%

Note: Each column denotes a different quality tier based upon the estimates of the finite mixture model. All businesses are weighted using the sampling weights. "Share High Risk" is the share of high risk businesses, as defined by the BBB.

to low quality business; the average rating falls from 3.9 for high quality businesses to 3.4 for medium quality businesses to 2.4 for low quality businesses. Ratings on Google, Facebook, and HomeAdvisor all decline much less when business quality falls; Google and Facebook's ratings falls by about 0.9 stars on average, and HomeAdvisor's ratings by 0.5 stars, when going from a high quality to low quality business. A low quality business on Google has about the same average rating as a medium quality business on Yelp or a high quality business on the BBB's platform.

Figure 4b depicts the estimates from panel regressions controlling for business fixed effects; as in Section 3, these results are relative to the BBB's rating. In Appendix C.1, I show that these findings are robust to using balanced panels with platform ratings for the BBB, Yelp, and Google for all businesses, or platform ratings for the BBB, Yelp, Google, and Facebook for all businesses.

Both high quality and low quality businesses on Yelp are about a half star higher rating than the BBB, while medium quality businesses on Yelp are about 1.2 stars higher than the BBB. Thus, the difference between Yelp and BBB ratings is similar for high quality and low quality businesses. Estimates for Google, Facebook, and HomeAdvisor are quite similar to



(a) Average Rating

(b) Rating Relative to BBB

Figure 4 Rating by Platform and Quality Tier

Note: Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights.

each other; the gap between their ratings and the BBB is higher for low quality businesses than high quality businesses. For low quality businesses, ratings on those three platforms are 1.7 stars higher on average relative to the BBB's platform.

4.3 Relative Rankings

The analysis so far has focused on differences in the average star rating between different quality businesses. However, platforms often use star ratings to rank businesses in search results to consumers; it is unclear whether these rankings would be affected. I examine whether higher quality businesses would be ranked higher through a simulation exercise. For each platform, I randomly draw a high quality business and low quality business and then record whether the rating for the high quality business was less than the low quality

Table VI Simulation Probabilities of Quality Rankings by Platform

Platform	High $<$ Low	High $<$ Medium	Medium $<$ Low
BBB	6.0	14.0	37.3
Yelp	9.0	28.7	23.7
Google	9.6	17.1	37.2
Facebook	9.4	25.9	27.6
HomeAdvisor	16.4	36.9	26.6

Note: All results reflect simulation estimates from one million simulations. For the second column, "High $<$ Low", each simulation randomly draws a high quality business and low quality business for each platform. Reported probabilities are averages across simulations of whether the high quality business is rated less than the low quality business. "High $<$ Medium" and "Medium $<$ Low" columns are defined analogously.

business. I estimate the probability that a high quality business has a lower rating than a low quality business by averaging across one million simulations.

High quality businesses are rarely rated below low quality businesses; Table VI includes the results of these simulations. A high quality business is rated greater than or equal to a low quality business 6% of the time for the BBB, 9% of the time for Yelp and Facebook, 10% of the time for Google, and 16% of the time for HomeAdvisor.

I then conduct the same exercise comparing high and medium quality businesses as well as medium and low quality businesses. Platforms that are more likely to rate a high quality business below a medium quality business tend to be relatively less likely to rate a medium quality business below a low quality business. The BBB and Google are the least likely to rate a high quality business below a medium quality business, but the most likely to rate a medium quality business below a low quality business.

4.4 Alternative Mixture Models

For my main results, I have estimated a 10 signal mixture model with 3 types. I have examined estimating a larger number of types; with four components, the medium quality tier is separated into two types. With five components, both the medium and high quality tiers are each separated into two types. However, the low quality type tends to be consistent when adding additional components. Much of the consumer protection interest is in identifying bad businesses, which in the model correspond to the low quality type. For example, consumer protection organizations might want to warn consumers about the low quality businesses or investigate them further. Thus, I use three tiers for my main analysis.

I have also examined only including the consumer protection measures as signals, and excluding the review ratings. I provide further infromation about this mixture model in Appendix C.2. Only including consumer protection measures as signals produces quality tiers that are similar to the BBB letter grade – the high quality type is mostly A+ graded businesses, the low type mostly F graded businesses, and the medium quality type with grades in between A+ and F. This is intuitive – using the information that the BBB has, the model replicates the BBB letter grade. By adding review signals, the mixture model identifies businesses that have a high BBB grade, but many complaints and low platform ratings, as low quality.

5 Fake Reviews as an Explanation

In this section, I first show that Google, Facebook, and HomeAdvisor tend to have more reviews than the BBB or Yelp, and low quality businesses have more reviews in general. I then use information from the filtering algorithms of the BBB and Yelp to assess whether fake reviews can account for these differences. I also examine differences in review length between platform, which may proxy for the quality of the review, using large review datasets from Google and Yelp. Finally, I develop a decomposition to assess quantitatively how much fake reviews contribute to the difference in average rating between Google and Yelp for low quality businesses.

5.1 Number of Customer Reviews

More fake reviews would increase the number of reviews on a platform. The number of reviews could also vary for other reasons, however. First, platforms may vary in their user base; consumers that are less likely to review may also have less extreme views, which could raise or lower average ratings. Dai et al. (2018) show how adjusting review ratings for differences across reviewers can affect average ratings. Second, the platform's policies may affect the number of reviews. Fradkin et al. (2017) study the effects of providing incentive payments for reviews on AirBnB, and find that incentive payments increased the number of reviews and number of negative reviews for both hosts and guests, although they decreased the number of one star reviews. Marinescu et al. (2018) examine Glassdoor and find that incentivising reviews – by requiring a review in order to view information on Glassdoor – induced more negative reviews.

In Table VII, I report the average number of customer reviews by platform and quality tier. For Facebook, I report an imputed measure using a coded sample because the number of customer reviews provided by the Google Search API tends to underestimate the true number of reviews.[19]

[19] A research assistant coded the number of reviews listed on Facebook pages to compare, and estimated a regression model of the number of reviews from Facebook pages on the number of reviews reported by the Google Search API. The regression model had an R^2 of 99.2%. I then use the regression model to impute the true number of Facebook reviews based upon the number of Facebook reviews listed by the Google Search API. The model is 7.7 + 1.19 * X where X is the number of Facebook reviews reported by the Google Search API, based on 344 observations.

Google, Facebook, and HomeAdvisor have more reviews, on average, than the BBB or Yelp. For example, the average high quality business on the BBB has 4 reviews, on Yelp 33, on Google 89, on Facebook 42, and on HomeAdvisor 49. In addition, low quality businesses tend to have more reviews on all platforms. The average low quality business on the BBB has 12 reviews, on Yelp 77, on Google 338, on Facebook 113, and on HomeAdvisor 101.

These results are consistent with an explanation of fake reviews increasing the number of reviews on Google, Facebook, and HomeAdvisor, especially for low quality businesses. On the other hand, they could also simply reflect differences in a user base across platforms. In addition, customers of low quality businesses could be more likely to post a review online, perhaps because consumers with "extreme" experiences are more likely to review.

Table VII Number of Reviews by Platform and Quality Tier

	BBB	Yelp	Google	Facebook	HomeAdvisor
High	3.9	32.8	88.9	41.9	49.3
	(0.0)	(1.5)	(2.9)	(1.8)	(2.5)
Medium	1.7	57.6	135.8	63.7	45.9
	(0.0)	(1.5)	(3.3)	(2.8)	(3.9)
Low	11.7	76.9	338.4	113.1	100.6
	(0.4)	(2.4)	(10.0)	(13.3)	(10.6)
Observations	59275	25792	60193	11240	2757

Note: Estimates clustered at the individual business level and weighted using the sampling weights. Number of Facebook reviews imputed as defined in the text.

5.2 Review Filter Algorithms

In order to further examine how fake reviews might affect review ratings, I use data from the review filtering algorithms of both the BBB and Yelp.

For Yelp, I have data on the reviews that were filtered by Yelp's proprietary algorithm for detecting fake reviews, and then hidden on Yelp's website. Luca and Zervas (2016) have provided evidence that these hidden reviews are a good proxy for fake reviews. In addition, the BBB has provided me data on the BBB's own proprietary filtering algorithm's score for each review, including both published and unpublished reviews. This score is meant to predict the probability that a review is fake.[20] I use the estimates of this algorithm to separate reviews into those with a very low probability of being fake and those with a high probability, or very high probability, of being fake.

Fake reviews could affect how ratings reflect business quality in two ways. First, the difference in rating between fake reviews and real reviews could be larger for low quality businesses. Second, fake reviews could be a larger share of reviews of low quality businesses.

I examine both of these channels using data from the review filtering algorithms. First, I examine whether low quality businesses are more likely to have reviews that are likely to be fake. Table VIII displays the share of BBB reviews deemed very likely to be fake, and Yelp reviews that are hidden, by quality tier.

Surprisingly, the share of reviews that are likely fake does not vary much across quality tiers. For the BBB, 7.9% of high quality businesses have an algorithm score identifying them as very likely to be fake, compared to 9.4% of low quality businesses. For Yelp, 46.6% of reviews of high quality businesses are hidden, compared to 46.0% of low quality businesses. One interpretation of these results is that all businesses have fake reviews – it is only low quality businesses for which these fake reviews affect the average rating. Another interpretation is that creators of fake reviews are good enough at spoofing real reviews that many legitimate reviews of high quality firms are flagged as fake.

[20] Reviews could not be published for reasons other than fake reviews, such as profanity, spam, or duplication of existing reviews or complaints, so the fake review score provides a more accurate guide to likely fake reviews.

Table VIII Share of Likely Fake Reviews by Quality Tier

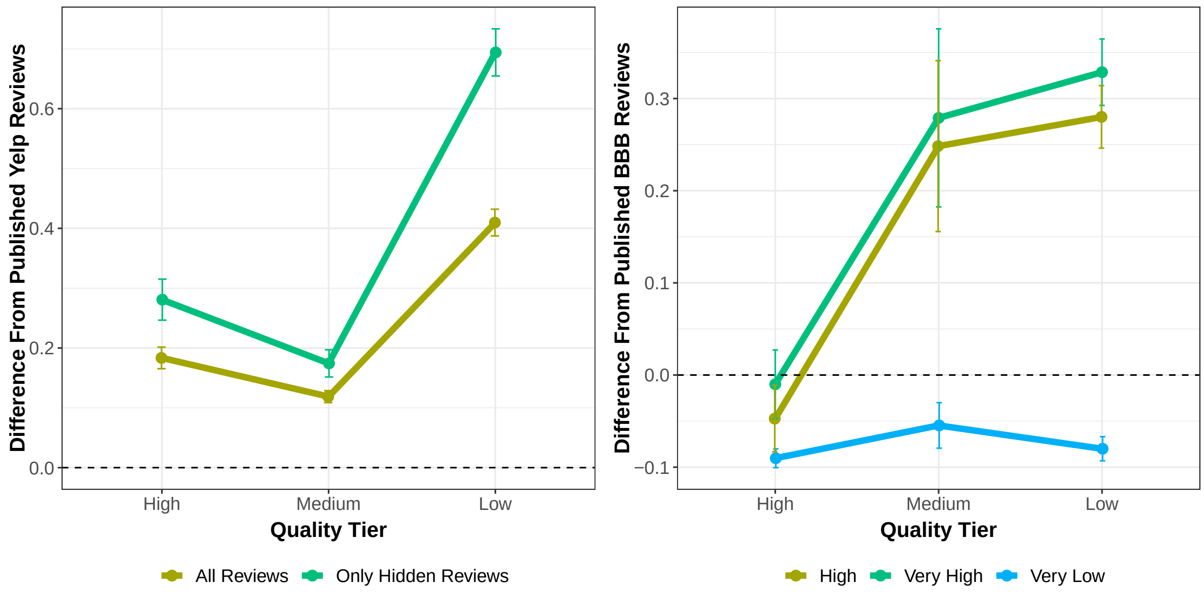
	BBB Very Likely to Be Fake	Yelp Hidden
High	7.9	46.6
	(0.2)	(0.4)
Medium	8.9	33.1
	(0.2)	(0.2)
Low	9.4	46.0
	(0.2)	(0.4)
Observations	72257	25792

Note: Estimates clustered at the individual business level and weighted using the sampling weights.

Next, I measure the difference in rating by quality tier between published reviews and likely fake reviews. To do so, I estimate panel regressions that control for individual business fixed effects, comparing alternative ratings for a platform to published ratings for different quality tiers as estimated in Section 4.

For Yelp, I examine ratings using all reviews – hidden and published – as well as only hidden reviews. I depict these results in Figure 5a. Including the filtered reviews would increase review ratings, especially for low quality businesses; the average rating would be 0.2 stars higher for high quality businesses and 0.1 stars higher for medium quality businesses, compared to 0.4 stars higher for low quality businesses. The average filtered ratings are 0.3 and 0.2 stars higher for high quality and medium quality businesses, compared to 0.7 stars higher for low quality businesses.

I adopt the same approach for the BBB. Review ratings based on the fake review probability from the filtering algorithm have a similar pattern to the Yelp results; Figure 5b displays these results. Reviews with a very low probability of being fake have slightly lower ratings – between 0.05 and 0.1 stars – than those published. High quality businesses have similar ratings using only ratings with high or very high probabilities of being fake. In con-



(a) Yelp Publication Status (b) BBB Fake Review Probability

Figure 5 Differences in Rating for Likely Fake Reviews for each Quality Tier

Note: All estimates relative to published ratings on Yelp (left figure) or the BBB (right figure). Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights.

trast, medium and low quality businesses have 0.25 to 0.35 stars higher scores when using ratings that are high or very high probability fake according to the algorithm.

Thus, reviews that are likely to be fake have higher ratings for low quality businesses for both BBB and Yelp reviews. However, the increase in the average rating of low quality businesses for reviews likely to be fake is only 0.7 stars on Yelp, and 0.3 stars for the BBB. These differences are much smaller than the gap in rating between Google, Facebook, and HomeAdvisor listings, on the one hand, and BBB and Yelp listings, on the other hand, for low quality businesses documented in Section 4.

To show this more fully, Figure 6 compares the average rating for BBB reviews predicted to be highly likely to be fake and hidden Yelp reviews to average Google, Facebook, and HomeAdvisor ratings for each quality tier. Google, Facebook, and HomeAdvisor continue to

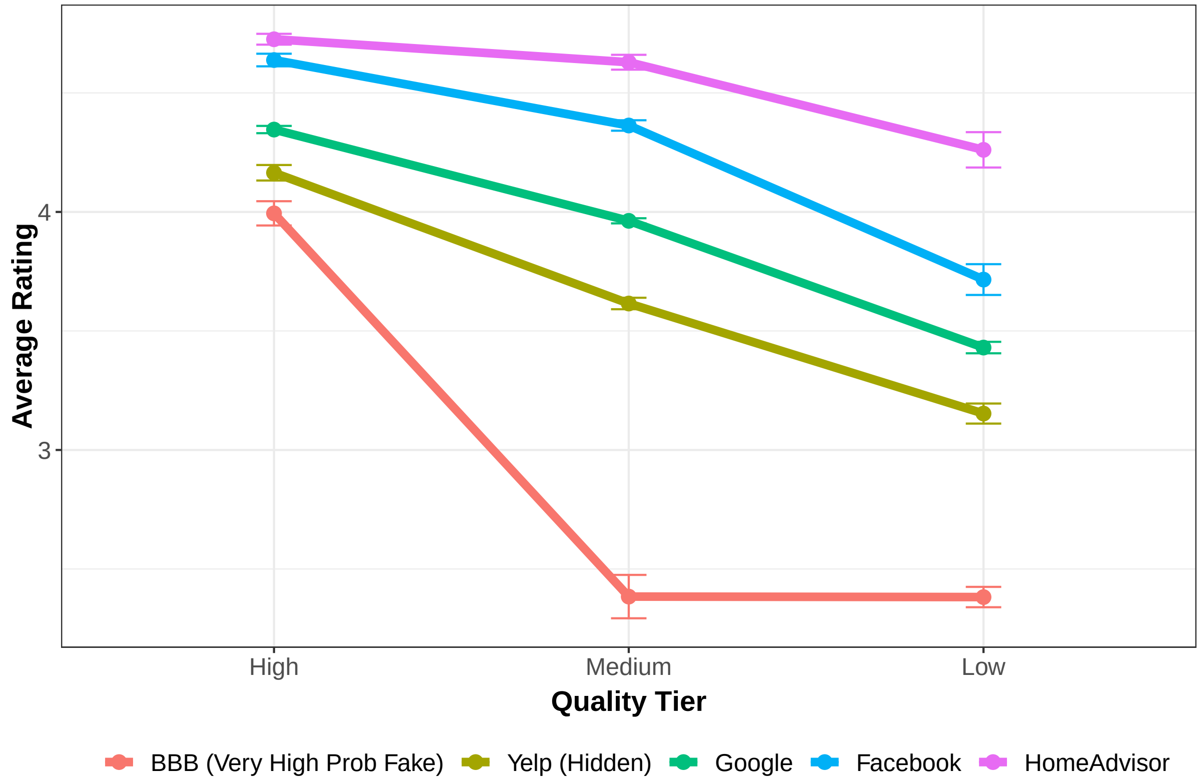


Figure 6 Average Rating for each Quality Tier Using Likely Fake BBB and Yelp Reviews

Note: All estimates relative to published BBB ratings. Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights.

have higher average ratings at every quality tier than BBB and Yelp reviews that are likely to be fake. However, the gap between these ratings shrinks for businesses in the low quality tier. Likely fake BBB and Yelp ratings are about 1 star and 0.3 stars lower than Google ratings, respectively, compared to 1.7 and 1 stars lower than Google ratings for published reviews for the same platforms.

5.3 Review Length

In this section, I examine an alternative explanation to fake reviews that could explain differences across platforms: that differences in review ratings across platforms reflect the type of review that users on those platforms write. I focus on the length of reviews.

Platforms vary substantially in their policies on review length, and have changed these policies over time. Yelp and the BBB do not allow ratings without any review text, while Google does allow such "no-text" ratings. Facebook used to allow no-text ratings, but now imposes a 25 character limit. With product reviews, Amazon has moved in the opposite direction by recently allowing no-text ratings.

Requiring a reviewer to write text imposes greater costs on reviewers, which might reduce the quantity of reviews but increase their quality. If higher quality reviews are, on average, more critical, requiring review text could lower average review ratings.

I examine this question using two auxiliary datasets that are large corpuses of reviews, as I only have individual reviews for businesses in my main dataset for the BBB and Yelp. For Google, I have data on 5.5 million reviews of US businesses collected by He et al. (2017) and Pasricha and McAuley (2018); most of these reviews are from 2010 to 2014. For Yelp, I use data from the Yelp Challenge, which contains 5.6 million reviews of US businesses; most of these reviews are from 2015 to 2018.[21] For both companies, I measure review length as the number of characters in the review. In this data, Google review ratings are, on average, 0.3 stars higher than Yelp review ratings, with an average of 4.05 for Google and 3.74 for Yelp.

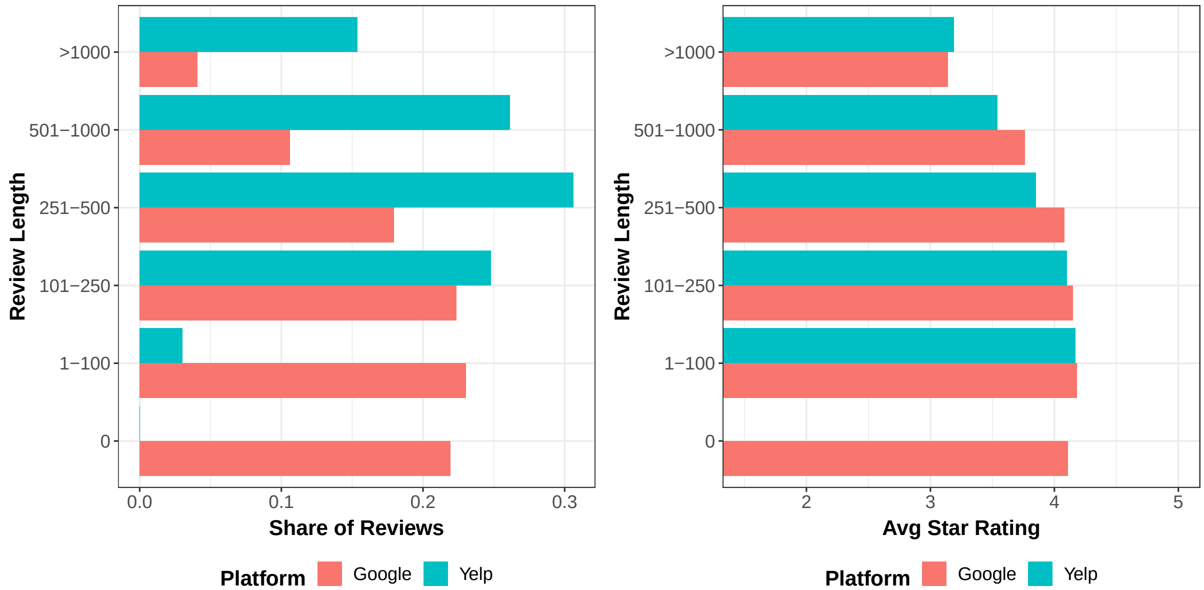
Reviews on Yelp are, on average, much longer than those on Google. The average review length on Yelp is 593 characters, more than double the average review count of 250 characters for Google. Figure 7a displays the share of reviews on both platform by review length

[21] The Google review data is available at https://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local. The Yelp challenge data is available at https://www.yelp.com/dataset; US reviews in the Yelp challenge data are from the Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland metropolitan areas.

category. For Google, 22% of reviews have no text. In addition, 23% of reviews on Google have between 1 and 100 characters, compared to 3% of reviews on Yelp with less than 100 characters. On the other hand, 11% of reviews on Google have 501 to 1,000 characters, and 4% have more than 1,000 characters, compared to 26% of Yelp reviews with 501 to 1,000 characters and 15% with more than 1,000 characters.

For both Google and Yelp, longer reviews have, on average, lower ratings. Figure 7b displays the average star rating on both platforms by review length category. No-text reviews on Google have an average of 4.1 stars; reviews with 1 to 100 characters have an average of 4.2 stars on both Google and Yelp. In contrast, reviews with 501 to 1,000 characters have 3.8 stars on Google and 3.5 stars on Yelp, and reviews with more than 1,000 characters have an average of 3.1 stars on Google and 3.2 stars on Yelp.

I then examine whether differences in review length can account for differences in the average rating across platforms through two simple counterfactual exercises. In these exercises, I hold constant the average rating by review length but change the distribution of review lengths across platforms. If Google had the average share of reviews by review length category as Yelp, its average review rating would be 3.87 (or 0.18 stars lower). The change in the review length distribution could then explain 57% of the difference in average rating between Google and Yelp. If Yelp had the average share of reviews by review length category as Google, its average review rating would be 3.99 (or 0.25 stars higher). The change in the review length distribution could then explain 81% of the difference in average rating between Google and Yelp. Thus, differences in review length have the potential to explain some of the differences between Google and Yelp.



(a) Share of Reviews (b) Average Rating

Figure 7 Share of Reviews and Average Rating by Review Length and Platform

Note: Reviews are reviews for US businesses from two large corpuses of reviews; for Google, collected by He et al. (2017) and Pasricha and McAuley (2018), and for Yelp, from the Yelp challenge dataset. See the text for further details.

5.4 Decomposition

So far, I have shown evidence that fake reviews likely explain some of the differences in ratings across platforms, but platforms differ for other reasons as well. We can make more quantitative statements on how fake reviews affect ratings by placing additional assumptions on the problem. In this section, I do so by assuming linearity and quality-invariant differences across platforms. These assumptions allow a "difference in difference" type analysis comparing average ratings for high and low quality businesses across platforms to identify the contribution from fake reviews to platform differences.

I build a structural framework in which platforms vary in ratings due to quality-invariant differences across platforms as well as fake reviews. The average rating Y_{ipq} is:

$$Y_{ipq} = \alpha_p + \gamma_{pq}\beta_q^F + (1 - \gamma_{pq})\beta_q^R + \epsilon_{ipq} \quad (5)$$

Here, Y_{ipq} is the average rating for business i on platform p with quality tier q, and is the sum of four terms. The first term is a platform fixed effect α_p, which captures systematic differences across platforms that are invariant to business quality. For example, platforms might have different types of users or different policies on posting reviews.

The probability that a review is fake for platform p and quality tier q is γ_{pq}, while β_q^F is a quality tier effect for fake (F) reviews and β_q^R is a quality tier effect for real (R) reviews. Thus, the second term is the contribution of fake reviews to the average rating – the probability of being fake γ_{pq} multiplied by the fake review quality tier effect β_q^F. The third term is the contribution of real reviews to the average rating – the probability of being real $(1 - \gamma_{pq})$ multiplied by the real review quality tier effect β_q^R. Finally, ϵ_{ipq} is an idiosyncratic error term that I assume is mean zero.

While this framework is fairly simple, it is underidentified without data on which reviews are fake. I examine a decomposition with two quality tiers, high and low, and two platforms, Google and Yelp. In that case, there are four moments – averages across businesses of ratings by platform and quality tier – but ten parameters: α_{Google}, α_{Yelp}, β_{high}^F, β_{high}^R, β_{low}^F, β_{low}^R, $\gamma_{Yelp,high}$, $\gamma_{Yelp,low}$, $\gamma_{Google,high}$, and $\gamma_{Google,low}$.

Given equation (5), the difference between the expected Google rating and the expected

Yelp rating for low quality businesses is:

$$\begin{aligned} E[Y_{i,Google,low}] - E[Y_{i,Yelp,low}] &= \\ (\alpha_{Google} - \alpha_{Yelp}) + (\gamma_{Google,low} - \gamma_{Yelp,low})(\beta^F_{low} - \beta^R_{low}) & \end{aligned} \tag{6}$$

Here, the difference between the expected Google rating and expected Yelp rating for low quality businesses is due to two factors: quality-invariant differences across platforms, represented by the difference in platform fixed effects $\alpha_{Google} - \alpha_{Yelp}$, and fake reviews, represented by the difference in fake review shares $\gamma_{Google,low} - \gamma_{Yelp,low}$ multiplied by the difference in average rating for fake reviews of low quality businesses compared to real reviews of those businesses $\beta^F_{low} - \beta^R_{low}$. Figure 4b showed that, after controlling for business fixed effects, low quality businesses have a 1.16 higher average star rating on Google compared to Yelp.

In order to make further progress, I assume that fake reviews do not affect the difference in average rating between Google and Yelp for high quality businesses. For this to be true, either the share of fake reviews could be the same for both platforms for high quality businesses (so $\gamma_{Google,high} = \gamma_{Yelp,high}$), or fake and real reviews for high quality businesses could have the same average rating ($\beta^F_{high} = \beta^R_{high}$). The latter assumption is broadly consistent with the evidence from the BBB and Yelp filtering algorithms presented in the previous section.

Given that assumption, the difference in average rating between Google and Yelp for high quality businesses identifies $\alpha_{Google} - \alpha_{Yelp}$. This is a "difference in difference" type of analysis; the decomposition compares ratings for high and low quality businesses across Google and Yelp, and the double difference is the estimate of the fake review effect. On average, Google ratings of high quality businesses are 0.55 stars above Yelp ratings of such businesses. Thus, 53% of the the difference between the expected Google rating and expected Yelp rating for low quality businesses is due to fake reviews, as opposed to 47% (0.55/1.16) due to quality-invariant differences across platforms.[22]

I cannot separately identify the difference in fake review shares $\gamma_{Google,low} - \gamma_{Yelp,low}$ from the difference in average rating of fake reviews compared to real reviews $\beta^F_{low} - \beta^R_{low}$. However, I can provide a lower bound for Google's fake review share. I use the fact that the average of real reviews on Google for low quality businesses should be higher than the average of real reviews on Yelp for such businesses given the estimated quality-invariant platform effects, and the average of fake reviews is bounded above by 5. In that case, because the average rating for low quality businesses on Yelp is 2.44, at least 30% of Google reviews of low quality businesses are fake. Appendix B.2 provides the proof of this lower bound.

So far, I have not used the data on likely fake reviews flagged by Yelp's algorithm to identify any parameters. One potential approach for estimation would be to use the empirical average rating difference between hidden and published reviews on Yelp for low quality businesses to estimate $\beta^F_{low} - \beta^R_{low}$. On average, hidden reviews on Yelp are 0.7 stars higher than published reviews for low quality businesses. If $\beta^F_{low} - \beta^R_{low} = 0.7$, at least 89% of Google reviews of low quality businesses would be fake.

[22]If fake reviews also increase Google's average ratings for high quality businesses compared to Yelp, then this approach would overstate quality-invariant platform differences, and 53% would be a lower bound on the contribution of fake reviews to the rating difference for low quality businesses.

6 Discussion and Conclusion

In this article, I have examined how review ratings vary across several different platforms, and shown that platforms differ systematically in their review ratings. Review ratings are much higher on Google, Facebook, and HomeAdvisor compared to the BBB and Yelp. After estimating a finite mixture model to place businesses into quality tiers, the gap in average rating between these platforms is the highest for low quality businesses.

I also found evidence that fake reviews explain some of these differences. The number of reviews is significantly higher on Google, Facebook, and HomeAdvisor compared to the BBB and Yelp. In addition, ratings for reviews that are more likely to be fake – reviews that are hidden on Yelp, and with high scores on an internal filtering algorithm for the BBB – are substantially higher than published reviews for low quality businesses. Through a decomposition, I estimated that about half of the difference between Google and Yelp ratings of low quality businesses are due to fake reviews, and at least about a quarter of Google reviews for low quality businesses are likely fake.

This article provides guidance to consumers, platforms, and regulators. For consumers, this research has shown that relying on the level of a business's star rating may not provide a good guide to business quality. The same 4.0 rating could imply a very different level of quality on one platform compared to another. On the other hand, the relative ranking of a business on a platform does appear to be more consistent across platforms. This may require consumers to search more in order to learn the distribution of ratings for a particular type of business and platform.

For platforms, this research has shown that a platform's policies and design choices, such as its algorithms to filter for fake reviews and its required review length, can substantially affect the ratings that businesses receive. A stronger filter for fake reviews will likely reduce average ratings for low quality businesses, for example. In addition, policies that increase the quality of reviews may decrease the quantity of reviews; platforms may need to communicate review quality to users through statistics beyond the number of reviews.

For consumer protection authorities, this research has shown that policing fake reviews is valuable, as fake reviews disproportionately boost ratings of low quality business with consumer protection problems. Second, given the documented differences in ratings across platforms, regulators could require platforms to provide information on the distribution of review ratings in order to assist consumers given that distribution varies so much across platforms. Lastly, the finite mixture model approach used in this paper may be helpful for consumer protection organizations to develop new quality measures for businesses, evaluate existing measures, and evaluate platform conduct.

For future research, it would be helpful to examine directly how reforms to platform conduct affect the distribution of reviews. For example, if a platform institutes a stricter filtering policy, does the share of fake reviews fall in the long run, or do fake reviews become more sophisticated? In addition, we know very little about what consumers believe about how reviews vary across different platforms, and whether their expectations match reality. Finally, it would be helpful to understand more how, and why, the characteristics of reviewers varies across platforms.[23]

[23]For example, Raval (2020) documents substantial selection in consumers who choose to complain, with victims in Black and Hispanic areas much less likely to complain about fraud.

References

Adams, Christopher P, "Finite Mixture Models with One Exclusion Restriction," *The Econometrics Journal*, 2016, *19* (2), 150–165.

Allman, Elizabeth S, Catherine Matias, and John A Rhodes, "Identifiability of Parameters in Latent Structure Models with Many Observed Variables," *The Annals of Statistics*, 2009, *37* (6A), 3099–3132.

Anderson, Eric T and Duncan I Simester, "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research*, 2014, *51* (3), 249–269.

Balan, David J, Patrick DeGraba, Francine Lafontaine, Patrick McAlvanah, Devesh Raval, and David Schmidt, "Economics at the FTC: Fraud, Mergers and Exclusion," *Review of Industrial Organization*, 2015, *47* (4), 371–398.

Benaglia, Tatiana, Didier Chauveau, and David R Hunter, "An EM-like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures," *Journal of Computational and Graphical Statistics*, 2009, *18* (2), 505–526.

_ , _ , David Hunter, and Derek Young, "mixtools: An R Package for Analyzing Finite Mixture Models," 2009.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin, "Non-parametric Estimation of Finite Mixtures from Repeated Measurements," 2016.

Cabral, Luis and Ali Hortacsu, "The Dynamics of Seller Reputation: Evidence From eBay," *The Journal of Industrial Economics*, 2010, *58* (1), 54–78.

Chauveau, Didier and Vy Thuy Lynh Hoang, "Nonparametric Mixture Models with Conditionally Independent Multivariate Component Densities," *Computational Statistics & Data Analysis*, 2016, *103*, 1–16.

Chevalier, Judith A. and Dina Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 2006, *43* (3), 345–354.

Dai, Weijia, Ginger Jin, Jungmin Lee, and Michael Luca, "Aggregation of Consumer Ratings: an Application to Yelp. com," *Quantitative Marketing and Economics*, 2018, *16* (3), 289–339.

Dellarocas, Chrysanthos, "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science*, 2006, *52* (10), 1577–1593.

Fradkin, Andrey, Elena Grewal, and David Holtz, "The Determinants of Online Review Informativeness: Evidence from Field Experiments on Airbnb," Technical Report, Working Paper 2017.

Hall, Peter, Amnon Neeman, Reza Pakyari, and Ryan Elmore, "Nonparametric Inference in Multivariate Mixtures," *Biometrika*, 2005, *92* (3), 667–678.

_ and Xiao-Hua Zhou, "Nonparametric Estimation of Component Distributions in a Multivariate Mixture," *The Annals of Statistics*, 2003, *31* (1), 201–224.

He, Ruining, Wang-Cheng Kang, and Julian McAuley, "Translation-based Recommendation," in "Proceedings of the Eleventh ACM Conference on Recommender Systems" 2017, pp. 161–169.

He, Sherry, Brett Hollenbeck, and Davide Proserpio, "The Market for Fake Reviews," *Available at SSRN*, 2020.

Jin, Ginger Zhe and Andrew Kato, "Price, Quality, and Reputation: Evidence from an Online Field Experiment," *The RAND Journal of Economics*, 2006, *37* (4), 983–1005.

_ **and Phillip Leslie**, "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards," *The Quarterly Journal of Economics*, 2003, pp. 409–451.

Kasahara, Hiroyuki and Katsumi Shimotsu, "Non-parametric Identification and Estimation of the Number of Components in Multivariate Mixtures," *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 2014, pp. 97–111.

Kruskal, Joseph B, "Three-way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics," *Linear Algebra and its Applications*, 1977, *18* (2), 95–138.

Kumar, Srijan and Neil Shah, "False Information on Web and Social Media: A Survey," *arXiv preprint arXiv:1804.08559*, 2018.

Langhe, Bart De, Philip M Fernbach, and Donald R Lichtenstein, "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 2016, *42* (6), 817–833.

Lappas, Theodoros, Gaurav Sabnis, and Georgios Valkanas, "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," *Information Systems Research*, 2016, *27* (4), 940–961.

Levine, Michael, David R Hunter, and Didier Chauveau, "Maximum Smoothed Likelihood for Multivariate Mixtures," *Biometrika*, 2011, pp. 403–416.

Lewis, Gregory and Georgios Zervas, "The Supply and Demand Effects of Review Platforms," *Unpublished manuscript*, 2020.

Luca, Michael, "Reviews, Reputation, and Revenue: The Case of Yelp. com," 2011.

_ **and Georgios Zervas**, "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 2016, *62* (12), 3412–3427.

Marinescu, Ioana, Nadav Klein, Andrew Chamberlain, and Morgan Smart, "Incentives Can Reduce Bias in Online Reviews," Technical Report, National Bureau of Economic Research 2018.

Mayzlin, Dina, Yaniv Dover, and Judith A. Chevalier, "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *The American Economic Review*, 2014, *104* (8), 2421–2455.

Murphy, Rosie, "Local Consumer Review Survey," 2019.

Pasricha, Rajiv and Julian McAuley, "Translation-based Factorization Machines for Sequential Recommendation," in "Proceedings of the 12th ACM Conference on Recommender Systems" 2018, pp. 63–71.

Plotkina, Daria, Andreas Munzel, and Jessie Pallud, "Illusions of TruthExperimental Insights into Human and Algorithmic Detections of Fake Online Reviews," *Journal of Business Research*, 2020, *109*, 511–523.

Raval, Devesh, "Which Communities Complain to Policymakers? Evidence from Consumer Sentinel," *Economic Inquiry*, 2019, *forthcoming.*

_ , "Whose Voice Do We Hear in the Marketplace?: Evidence from Consumer Complaining Behavior," *Marketing Science*, 2020, *39* (1), 168–187.

Rayana, Shebuti and Leman Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2015, pp. 985–994.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood, "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics*, 2006, *9* (2), 79–101.

Shehnepoor, Saeedreza, Mostafa Salehi, Reza Farahbakhsh, and Noel Crespi, "Netspam: A Network-based Spam Detection Framework for Reviews in Online Social Media," *IEEE Transactions on Information Forensics and Security*, 2017, *12* (7), 1585–1595.

Tadelis, Steven, "Reputation and Feedback Systems in Online Platform Markets," *Annual Review of Economics*, 2016, *8*, 321–340.

_ **and Florian Zettelmeyer**, "Information Disclosure as a Matching Mechanism: Theory and Evidence from a Field Experiment," *American Economic Review*, 2015, *105* (2), 886–905.

Wu, Guangyu, Derek Greene, Barry Smyth, and Pádraig Cunningham, "Distortion as a Validation Criterion in the Identification of Suspicious Reviews," in "Proceedings of the First Workshop on Social Media Analytics" 2010, pp. 10–13.

Ye, Junting, Santhosh Kumar, and Leman Akoglu, "Temporal Opinion Spam Detection by Multivariate Indicative Signals," *arXiv preprint arXiv:1603.01929*, 2016.

A Data Appendix

I clean the Google Search and Place API results by comparing the address, zip code, and name of the business in the API results to the same fields in the BBB Business Register. The type of data provided varies by platform, as described below:

1. For Yelp and HomeAdvisor listings, the Google Search API provides the business name, street address, city, state, zip code, and country name in separate fields in a standardized format.

2. For Google listings, the Google Places API provides the business name in one field, and the full address (street address, city, state, and zip code) in another field in a standardized format.

3. For Facebook listings, the Google Search API provides the business name in one field, as well as a "snippet" that typically contains the business name and full address together with a description of the business, and another field that provides the city and state. Thus, for Facebook listings, I have to separate the snippet into separate fields for street address, city, state, and zip code; some listings do not contain zip code or address information, and the snippet format varies considerably across listings, making matching more challenging than for Google, Yelp, or HomeAdvisor listings.

I first exclude all listings where the state does not match, as well as listings where the street address or zip code are missing. I then construct measures of whether the listing matches the BBB Business Register on three criteria: business name, first line of business street address (i.e. before the city, state, and zip code), and business zip code. I only include listings for which the name, street address, and zip code all match. I use the following matching criteria:

1. For the name match, I use the Jaro-Winkler distance with $p = 0.1$, and consider the name to have matched if the Jaro-Winkler distance between the BBB Register name and API name is less than or equal to 0.25.

2. For the street address match, I first use the Jaro-Winkler distance with $p = 0.1$, and consider the street address to have matched if the Jaro-Winkler distance between the BBB Register street address and API street address is less than or equal to 0.25. In addition, to make sure that addresses with a different house number are not considered a match, I also require the first four characters of the BBB register street address and API street address to have a Levenshtein distance of 1 or less if the first two characters of the street address is a number.

3. For the zip code match, I use whether the zip code in the BBB Register is the same as the API zip code.

As stated above, unlike Google, Yelp, or HomeAdvisor listings, Facebook listings often have varying formats for the address within a snippet containing the business name and other details. Thus, for Facebook, I also consider an address to have matched if the string of the first 10 characters of the address is contained within the snippet, and I also consider a zip code has having matched if the full zip code is contained within the API snippet. These rules allow matches when the address or zip code is contained within the snippet in a non-standard way.

In order to examine how well this matching process worked, a Research Assistant checked a 600 entry random sample for each platform by going to the platform website and verifying if the

Table A-1 Matching Accuracy

	Categories Matching		
Platform	All Three	Two	One
Google	99.5%	66.5%	24.0%
Yelp	99.5%	76.0%	29.5%
Facebook	96.9%	75.8%	28.3%
HomeAdvisor	99.0%	69.5%	26.5%

Note: The number of categories matched refers to matches on business name, business street address, and business zip code. For Google, Yelp, and HomeAdvisor, the number of observations for the estimate in each column is 200. For Facebook, the number of observations for the estimate in each column is lower because some Facebook pages are private and could not be accessed. For the column of all three categories matching, the sample size is 191.

business is the same. The random sample was stratified to equally split between three categories: a full match (on name, address, and zip code), a match on two of three categories, and a match on one of three categories. The Research Assistant was not informed about the match quality.

In Table A-1, I display estimates of matching accuracy using this random sample. Of the entries with a full match on all three categories, 99.5% of the Yelp entries, 99.5% of the Google entries, 96.9% of the Facebook entries, and 99.0% of HomeAdvisor entries are coded as correctly matching.[24]

I also find significant drop-offs in match quality when not all three categories match. For Google, entries for which only two of three categories match are 67% correct, and for which only one of the three match are 24% correct. Similarly, for Yelp, entries for which two of the three match are 76% correct, and one of the three match are 30% correct. For Facebook, entries for which two of the three match are 76% correct, and one of the three match are 28% correct. Finally, for HomeAdvisor, entries for which two of the three match are 70% correct, and one of the three match are 27% correct.

In a small number of cases for Yelp, HomeAdvisor, and Facebook, I have multiple entries for the same platform and the same business. Many of these are the same entry (with say an international website of the platform), but for Facebook in particular the business sometimes has multiple different pages. When there are multiple entries, I choose the entry with the maximum number of reviews, and, if multiple entries remain, on lowest search rank.

For Yelp, I match the cleaned entries to data on all reviews directly provided by Yelp – 308 entries do not match which I exclude. For my measure of the average star rating for Yelp, I use the average of all Yelp ratings provided by Yelp as the Google Search API provides the average rounded to the nearest 0.5 (as reported on Yelp's website).

[24] The Facebook result is only based on a sample of 191; some entries could not be coded as the Facebook pages were not available to all users (i.e. they were private).

B Theory Appendix

B.1 Levine et al. (2011) Algorithm

First, define the smoothing operator $\mathcal{N}$ as:

$$\mathcal{N}f(x) = \exp \int K_h(x-u)\log f(u)du$$

and

$$\mathcal{N}f_j(x_i) = \prod_{k=1}^{K} \mathcal{N}f_{jk}(x_{ik})$$

where K_h is a kernel density function with bandwidth h.

Start with initial guesses for the type shares λ^0 and signal distributions f^0. Then iterate for $t = 0, 1, ...$ over the majorization and minorization steps:

1. Majorization Step:
$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N} f_j^t(x_i)}{\sum_{a=1}^{J} \lambda_a \mathcal{N} f_a^t(x_i)}$$

2. Minorization Steps:
$$\lambda_j^{t+1} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^t$$
$$f_{jk}^{t+1}(u) = \frac{1}{nh\lambda_j^{t+1}} \sum_{i=1}^{n} w_{ij}^t K\left(\frac{u - x_{ik}}{h}\right)$$

I implement this algorithm using the R package mixtools (Benaglia et al., 2009b).

B.2 Lower Bound for Fake Review Rate

Given the assumption that either $\gamma_{Google,high} = \gamma_{Yelp,high}$ or $\beta_{high}^F = \beta_{high}^R$,

$$E[Y_{i,Google,high}] - E[Y_{i,Yelp,high}] = \alpha_{Google} - \alpha_{Yelp} > 0 \tag{7}$$

In that case, we must have that:

$$E[Y_{i,Google,low}^R] = \alpha_{Google} + \beta_{low}^R = E[Y_{i,Yelp,low}^R] + (\alpha_{Google} - \alpha_{Yelp}) \tag{8}$$

Since reviews are bounded above by 5, we have that:

$$E[Y_{i,Yelp,low}] = \gamma_{Yelp,low} E[Y_{i,Yelp,low}^F] + (1-\gamma_{Yelp,low})E[Y_{i,Yelp,low}^R] \tag{9}$$

$$E[Y_{i,Yelp,low}^R] = \frac{E[Y_{i,Yelp,low}] - \gamma_{Yelp,low}E[Y_{i,Yelp,low}^F]}{1-\gamma_{Yelp,low}} \tag{10}$$

$$E[Y_{i,Yelp,low}^R] \geq \frac{E[Y_{i,Yelp,low}] - 5 * \gamma_{Yelp,low}}{1-\gamma_{Yelp,low}} \tag{11}$$

Thus, we have that:

$$E[Y^R_{i,Google,low}] = \alpha_{Google} + \beta^R_{low} = E[Y^R_{i,Yelp,low}] + (\alpha_{Google} - \alpha_{Yelp}) \quad (12)$$

$$E[Y^R_{i,Google,low}] > (\alpha_{Google} - \alpha_{Yelp}) + \frac{E[Y_{i,Yelp,low}] - 5 * \gamma_{Yelp,low}}{1 - \gamma_{Yelp,low}} \quad (13)$$

Now, since reviews are bounded above by 5, $E[Y^F_{i,Google,low}] = \alpha_{Google} + \beta^F_{low} \leq 5$. Thus,

$$E[Y^F_{i,Google,low}] - E[Y^R_{i,Google,low}] = \beta^F_{low} - \beta^R_{low} \leq 5 - (\alpha_{Google} - \alpha_{Yelp}) - \frac{E[Y_{i,Yelp,low}] - 5 * \gamma_{Yelp,low}}{1 - \gamma_{Yelp,low}} \quad (14)$$

Combining terms, we have that:

$$\beta^F_{low} - \beta^R_{low} \leq \frac{5 - (1 - \gamma_{Yelp,low})(\alpha_{Google} - \alpha_{Yelp}) - E[Y_{i,Yelp,low}]}{1 - \gamma_{Yelp,low}} \quad (15)$$

The fake review term in the decomposition was:

$$(\gamma_{Google,low} - \gamma_{Yelp,low})(\beta^F_{low} - \beta^R_{low}) = X \quad (16)$$

Substituting in the bound, we have:

$$(\gamma_{Google,low} - \gamma_{Yelp,low}) = \frac{X}{(\beta^F_{low} - \beta^R_{low})} \quad (17)$$

$$\gamma_{Google,low} \geq \frac{X(1 - \gamma_{Yelp,low})}{5 - (1 - \gamma_{Yelp,low})(\alpha_{Google} - \alpha_{Yelp}) - E[Y_{i,Yelp,low}]} + \gamma_{Yelp,low} \quad (18)$$

Using the values in the text, $X = 0.61$, $\alpha_{Google} - \alpha_{Yelp} = 0.55$, and $E[Y_{i,Yelp,low}] = 2.44$.

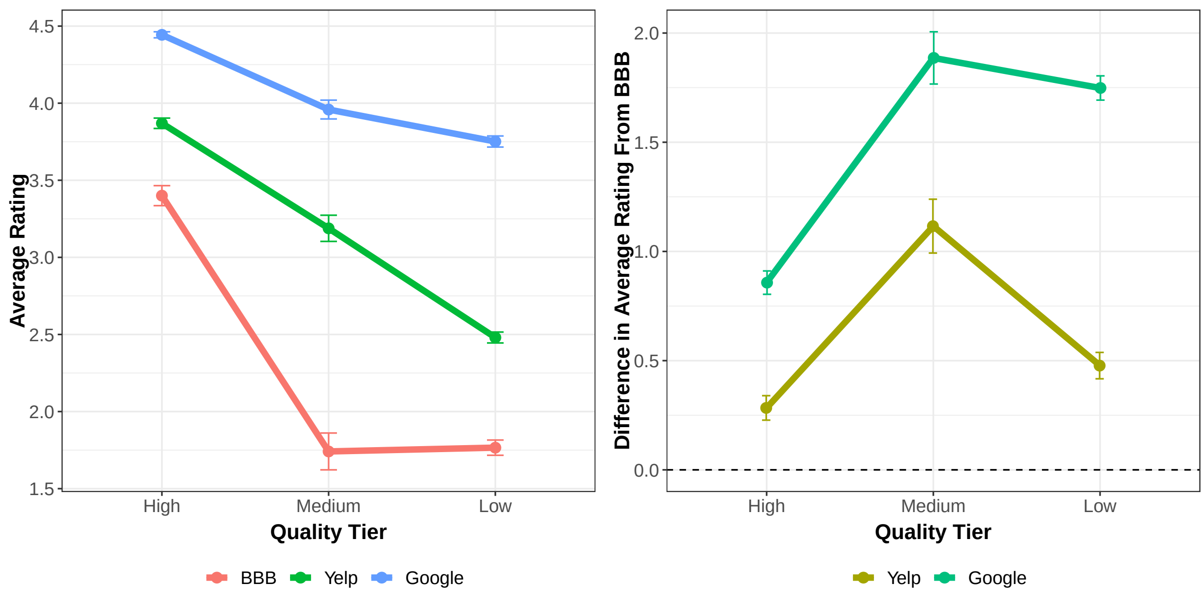
$$\gamma_{Google,low} \geq \frac{0.61(1 - \gamma_{Yelp,low})}{2.01 + 0.55\gamma_{Yelp,low}} + \gamma_{Yelp,low} \quad (19)$$

The lower bound for $\gamma_{Google,low}$ is when $\gamma_{Yelp,low} = 0$; in that case, $\gamma_{Google,low} = 0.30$.

C Empirical Appendix

C.1 Balanced Panels

In this section, I examine how average ratings vary by quality tier using balanced panels of either only businesses with BBB, Yelp and Google ratings (in Figure 8a and Figure 8b) or only businesses with BBB, Yelp, Google,and Facebook ratings (in Figure 9a and Figure 9b). Businesses with BBB, Yelp, and Google ratings comprise 6.2% of the sample, while businesses with BBB, Yelp, Google, and Facebook ratings comprise 1.2% of the sample. Estimates using both balanced panels are similar to using the overall sample.



(a) Average Rating

(b) Rating Relative to BBB

Figure 8 Rating by Platform and Quality Tier Using Balanced Panel with BBB, Yelp, and Google Ratings

Note: Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights.

C.2 Mixture Model Without Review Ratings

In this section, I examine estimates of a quality measure from a finite mixture model that only uses the five consumer protection signals, and does not use review ratings. Table A-2 provides summary statistics; this quality tiering follows the BBB letter grade closely. Figure 10a provides the average star rating by platform and quality tier. Figure 10b depicts the estimates from panel regressions controlling for business fixed effects; as in Section 3, these results are relative to the BBB's rating. Estimates using this quality tier are similar to those reported in the text using all 10 signals, except the difference between Yelp and Google for low quality businesses is smaller.

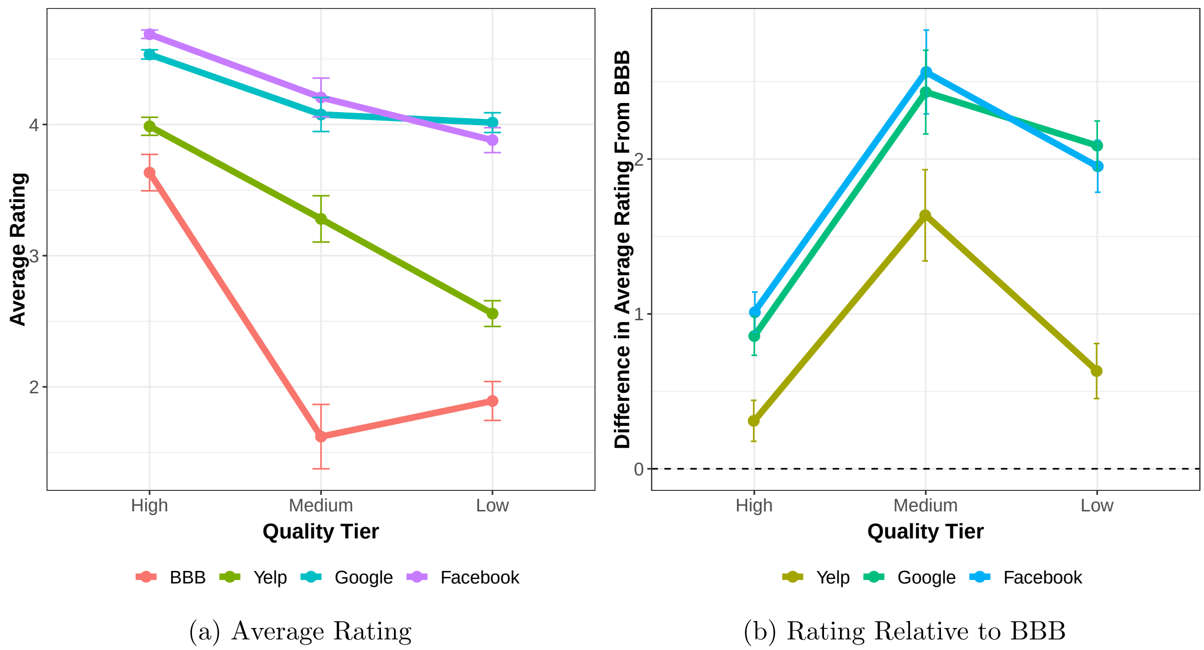


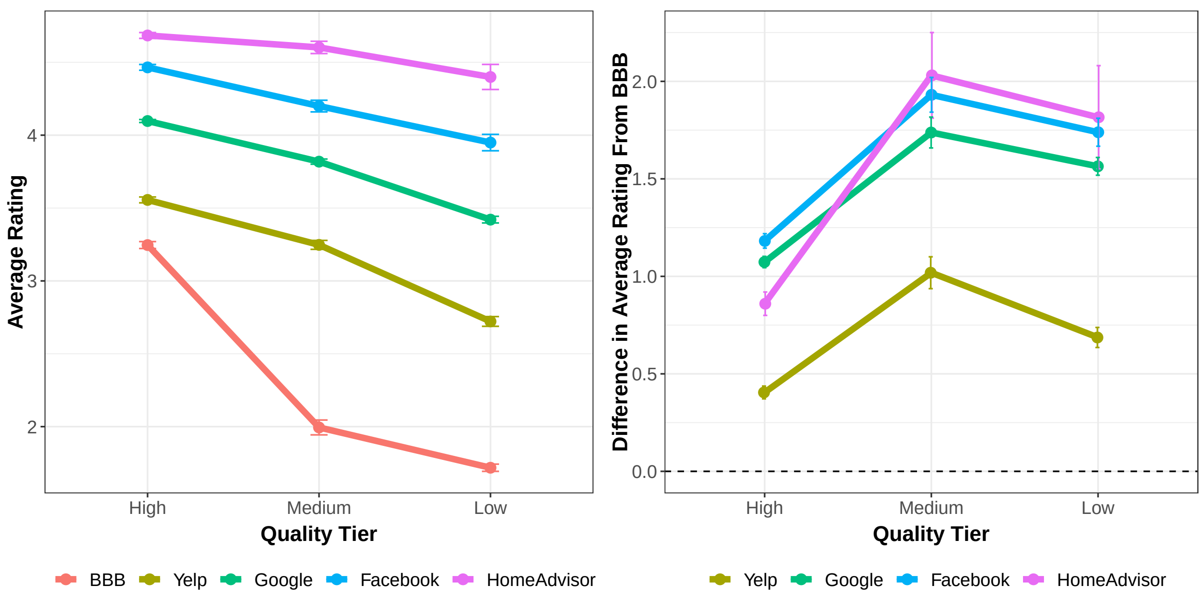
Figure 9 Rating by Platform and Quality Tier Using Balanced Panel with BBB, Yelp, Google, and Facebook Ratings

Note: Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights.

Table A-2 Summary Statistics by Quality Tier, for Quality Tiers Estimated Excluding Review Ratings

	High	Medium	Low
Type Share	70.4%	20.7%	8.9%
Median Number of Complaints	1.0	1.0	3.0
Mean Number of Complaints	2.2	1.7	8.6
Share High Risk	0.0%	0.1%	3.5%
Share With Complaints	72.6%	99.7%	98.3%
Share with A+ Grade	76.0%	0.0%	0.0%
Share with F Grade	0.0%	0.0%	81.4%
Share of Complaints Unresolved	5.6%	92.4%	79.8%

Note: Each column denotes a different quality tier based upon the estimates of the finite mixture model, where the finite mixture model only includes consumer protection signals and does not include review ratings. All businesses are weighted using the sampling weights. "Share High Risk" is the share of high risk businesses, as defined by the BBB.



(a) Average Rating

(b) Rating Relative to BBB

Figure 10 Rating by Platform and Quality Tier, for Quality Tiers Estimated Excluding Review Ratings

Note: Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights. Quality tiers estimated using a finite mixture model that only includes consumer protection signals and does not include review ratings.