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

Devesh Raval

Federal Trade Commission

draval@ftc.gov

May 23, 2024

Abstract

I examine how online reviews vary across competing online review platforms, examining both generalist platforms such as Google and Facebook as well as specialists such as the BBB and Yelp. I assess reviews across platforms by estimating quality tiers for local businesses through a finite mixture model using signals from consumer protection authorities and review ratings, and then comparing review ratings to the quality tiers. While review ratings are higher for Google and Facebook compared to the BBB and Yelp for all quality tiers, the gap between these platforms is substantially larger for low quality businesses with likely consumer protection problems. Finally, one dimension of platform quality is removing fake reviews; reviews that are likely fake based on multiple review filtering algorithms increase ratings for low quality businesses.

Keywords: online reviews, online platforms, 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. This paper was previously circulated as "Do Gatekeepers Develop Worse Products? Evidence from Online Review Platforms". 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, Ben Casner, Dan Hosken, Ginger Jin, Louis Kaplow, Patrick McAlvanah, Yesim Orhun, Jan Pappalardo, Rubens Pessanha, Ted Rosenbaum, Dave Schmidt, Kathy Spier, Andrew Sweeting, Charles Taragin, Will Violette, Mo Xiao, and Dan Wood 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, and by doing so provide signals about the quality of businesses and products to the marketplace. Consumers consult online reviews when making purchasing decisions, while platforms use online reviews to select how to rank products to display to consumers. In addition, online reviews play a crucial role by allowing businesses to gain reputations and thereby facilitating trust in the marketplace (Tadelis, 2016).

Policymakers, however, have been concerned with the quality of online reviews for both competition and consumer protection reasons. These concerns include gatekeeper platforms preferencing their own review services over rivals, as well as the prevalence of fake reviews. Poor quality reviews might lead customers to choose worse options in the marketplace, or rely on other signals of quality such as brands over reviews. Unfortunately, we know very little about how the quality of reviews varies across platforms.

I examine how online reviews compare across platforms by examining review listings from five platforms – the Better Business Bureau (BBB), Yelp, Google, Facebook, and HomeAdvisor. Consumer complaints to consumer protection authorities provide a signal of ground truth separate from the reviews hosted on these platforms, and so allow me to evaluate online reviews on different platforms.

In Section 2, I discuss the competition and consumer protection concerns concerning online reviews. On the competition side, the Federal Trade Commission (FTC) investigated Google for anticompetitive conduct in its preferencing its own vertical services, including

reviews, over independent competitors, as well as for scraping reviews from competitors.[1] The main potential harm cited from these actions was to reduce innovation and quality in these services.[2] Several US states recently sued Google over similar issues. On the consumer protection side, a major, persistent concern has been fake reviews. As part of its consumer protection mission, the FTC has investigated several firms and platforms for faking or suppressing reviews. It is now undertaking a major rulemaking banning fake reviews and testimonials.

Section 3 details how I match a sample of over one hundred thousand businesses to review listings on five platforms. Google and Facebook are dominant platforms in search and social media markets; reviews are a small part of their business. On the other hand, the BBB, Yelp, and HomeAdvisor all primarily focus on consumer reviews.

I then assess platforms by comparing review ratings with a measure of the quality of the business reviewed; a poor quality platform will continue to have high ratings for low quality businesses. I first show in Section 4 that business ratings on Google, Facebook, and HomeAdvisor tend to be significantly higher than on the BBB and Yelp. These differences are magnified for business with poor quality, as proxied by a low letter grade from the BBB or high numbers of consumer complaints.

These signals allow me to measure business quality in Section 5 by estimating a non-parametric finite mixture model. This model separates businesses into quality tiers based

[1]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.

[2]For example, a vertical provider states with respect to Google's conduct (Nadler and Cicilline, 2020) that "The anticompetitive effects reduce Google's own incentives to improve the quality of its services, because it does not need to compete on the merits with rival services."

upon signals of poor quality; I use signals from complaints to consumer protection organizations, such as a large number of complaints or a F grade from the BBB, as well as review ratings themselves. Because these measures of quality are based in part on consumer protection complaints, low quality businesses are likely to cause consumers harm.

I estimate three quality tiers that reflect the likelihood of experiencing consumer protection issues. 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.

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.

There are several potential explanations for differences in review quality across platforms, including differences in the composition of reviewers, varying reciprocity normes across platforms, and distortions from imperfect competition. I am able to empirically examine two potential explanations for differences in platform quality – fake reviews and platform rules on the type of content posted – in Section 6.

I examine fake reviews through proxies for whether a review is fake; whether the review is "hidden" from view on Yelp because it is flagged by Yelp's algorithms as likely fake (Luca and Zervas, 2016), and the score from the BBB's proprietary filtering algorithm predicting the probability that a review is fake. For both platforms, 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. By contrast, the difference for high quality businesses is much smaller for Yelp, and negligible for the BBB. Thus, a platform that spends less effort on reducing fake reviews is likely to have inflated reviews for low quality businesses.

Platforms also differ in the effort that consumers place in writing reviews. Unlike Yelp, Google allows "no-text" reviews. For my sample, 50% of Google reviews are 100 characters or less, many of which are no-text reviews, while only 2% of Yelp reviews are 100 characters or less. Thus, Google tends to have many more reviews but its reviews provide less information. Removing Google reviews with less than 100 characters reduces Google ratings for both high quality and low quality businesses, and so cannot explain the larger gap in rating between platforms for low quality businesses.

Researchers in economics and marketing have studied how signals of business quality affect 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. Luca (2011) and Lewis and Zervas (2020) find that a substantial increase in

restaurant revenue and hotel demand, respectively, with higher ratings on online platforms.

Only a few papers examine multiple review sites. 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. Similarly, Zervas et al. (2021) compare reviews on Airbnb to those on TripAdvisor, and find weak correlation between the two and some evidence of higher ratings on Airbnb. Fang (2022) examines restaurant reviews on Google and Yelp and finds, as I do, that review ratings tend to be higher on Google than Yelp.

Platforms have considerable latitude in how to measure and communicate business quality. Nosko and Tadelis (2015) show that buyers on eBay typically do not post negative reviews, and that a measure of seller quality based on the fraction of purchases with a review can help to promote higher quality sellers. Fradkin et al. (2021) and Fradkin and Holtz (forthcoming) show that negative experiences are underreported on AirBnB, and that either paying consumers to review or having sellers and buyers simultaneously review can reduce this underreporting. However, paying consumers to review did not affect sales and likely reduced welfare.

Finally, this paper is related to the recent literature in economics and marketing on fake reviews. Mayzlin et al. (2014) identify fake reviews through differences in platforms, comparing verified reviews on Expedia to unverified reviews on TripAdvisor. 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).[3]

[3]An extensive literature in computer science has also examined fake reviews (Kumar and Shah, 2018), fo-

2 Background

Low quality in review markets has been a persistent concern of regulators for both competition and consumer protection reasons. I discuss the history of both of these issues in detail below.

2.1 Competition and Gatekeeper Conduct

From 2010 to 2012, the FTC investigated Google for potential anticompetitive conduct, but decided against an enforcement action in the matter. Of the four main counts investigated, two related to rival vertical platforms that hosted local business reviews.

The first count concerned Google preferencing its own vertical properties and demoting rivals in Google search rankings. Through "Universal Search", Google began to place a box for Google Local with prominent visuals at the top of the search page when consumers searched for local businesses. Only Google properties were eligible for this placement, which pushed organic links to competitors such as Yelp or TripAdvisor farther down the page.

The main theory of harm was that the foreclosure of rivals would harm innovation. As the Bureau of Competition (BC) memo states[4]:

> The theory of harm to competition is mainly one of reduced innovation: that, when faced with Google's seamless ability to enter into highly monetizable categories of commerce and simultaneously to disadvantage its competitors, existing competitors cannot innovate at the same pace; new or innovative vertical websites

cusing 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).

[4]See https://stratechery.com/wp-content/uploads/2021/03/Staff-Memo.pdf for the BC staff memo.

> will cease to enter the market; and consumers will be faced with a corresponding reduction in innovation and choice.

Nevertheless, the BC memo argued against including the count in part because these design changes "have improved its [Google's] product by providing consumers with direct, relevant, and "better" results." Both the anti-competitive rationale for this count as well as Google's pro-competitive defense were based upon how Google's self-preferencing would affect the quality and innovation of online review platforms.

The second count concerned Google including reviews scraped from rivals as part of Google Local. Because online review sites with more reviews are more valuable to consumers, such scraping helped start up Google Local as a useful review site.[5] When rivals complained about this practice, Google threatened to cut them off from Google search results entirely if they did not acquiesce to review scraping.

The BC memo recommended that the Commission issue a complaint on this count, based on harms to innovation of both Google and its rivals:

> More broadly, the natural and probable effect of Google's conduct is to diminish the incentives of other vertical websites to invest in, and to develop, new and innovative products. Entrepreneurs may be reluctant to develop new websites, and investors may be reluctant to sponsor that development, recognizing that Google can use its monopoly power over search to simply appropriate competing content that it deems lucrative to its own search empire. Further, Google's conduct suggests that Google itself has failed to innovate, as it would have to have done in the absence of scraping content from its rivals.

While the FTC opted not to bring this lawsuit, 38 states sued Google in 2020 on antitrust grounds which included very similar issues of gatekeeper conduct that reduced traffic to

[5]Google eventually stopped this practice and removed the scraped reviews, although the scraped reviews gave it a head start to attract consumers. The BC memo states: "Google had already collected sufficient reviews by bootstrapping its review collection on the display of other websites' reviews. It no longer needed to display third-party reviews, particularly while under investigation for this precise conduct."

specialized vertical providers, including review sites.[6] The states' complaint stated:

> By eliminating competitive constraints in its search-related markets, Google has become a monopolistic gatekeeper, free to limit passage across the internet and to charge supracompetitive tolls for the journey. Through its anticompetitive conduct, Google has gained and maintained the power to redirect or choke-off the consumer traffic flowing to specialized vertical providers.

This lawsuit was later consolidated with the Department of Justice's lawsuit against Google for its behavior in search markets. In 2023, Judge Mehta dismissed the claims with respect to specialized vertical providers due to lack of evidence.[7]

2.2 Consumer Protection and Fake Reviews

Policymakers have expressed concern that consumers are exposed to fake or misleading reviews; most consumers now believe that they have read fake reviews online (Murphy, 2019).[8] In response, the FTC is using all of its weapons, including enforcement actions and rulemaking, to prevent such fake reviews.

The FTC has now brought several enforcement actions alleging review manipulation by a reviewed business or an online platform hosting reviews. The recent high profile *Sunday Riley* case provides an illustrative example; the FTC alleged that a company's CEO wrote, and ordered employees to write, five star reviews of the company's products on different platforms using false identities.[9] In addition, the FTC recently alleged in the *FashionNova*

[6] See https://portal.ct.gov/-/media/ag/press_releases/2019/02b---attachment-1---colorado-et-al-v-google-public-redacted-complaint.pdf for the complaint in this case.

[7] See https://www.documentcloud.org/documents/23897767-usa-v-google?responsive=1&title=1 for Judge Mehta's hearing at the summary judgment stage.

[8] 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.

[9] See https://www.ftc.gov/news-events/press-releases/2019/10/devumi-owner-ceo-settle-ftc-charges-they-sold-fake-indicators. Additional FTC cases on fake reviews include Cure Encap-

case that a *platform* suppressed hundreds of thousands of negative reviews of products on the platform as reviews below 4 stars were not shown to consumers.[10] The Competition and Markets Authority (CMA) of the UK has launched an investigation into Amazon and Google for potentially not doing enough to filter out fake reviews.[11]

The FTC has also used warning letters as a tool, recently sending warning letters to review management firms hired by websites to manage reviews. In October 2021, the FTC also sent a Notice of Penalty Offenses to over several hundred businesses warning them that they might be subject to civil penalties if they allow fake reviews or other deceptive endorsements on their websites.[12]

Finally, the FTC has begun a major rulemaking on the use of reviews and endorsements aimed at promoting transparency and protecting consumers from misleading information.[13] Violators of any eventual rules from this rulemaking would be potentially subject to civil

sulations, Urthbox, Mikey & Momo (aka Aromaflage), Universal City Nissan (aka Sage Auto), Son Le and Bao Le (aka Trampoline Safety of America), Amerifreight, and LendEDU. 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-\OT1\textquoteleftcustomer\OT1\textquoteright-reviews.

[10]FashionNova settled this case for $4.2 million. See https://www.ftc.gov/news-events/news/press-releases/2022/01/fashion-nova-will-pay-42-million-part-settlement-ftc-allegations-it-blocked-negative-reviews. The FTC has also brought a case against the Roomster platform for fake review activities; see https://www.ftc.gov/news-events/news/press-releases/2022/08/ftc-states-sue-rental-listing-platform-roomster-its-owners-duping-prospective-renters-fake-reviews.

[11]For the CMA investigation, see https://www.gov.uk/cma-cases/online-reviews.

[12]The Notice of Penalty Offenses allows the FTC to seek civil penalties against a company that engages in conduct that it knows has been found unlawful in a previous FTC administrative order. See https://www.ftc.gov/news-events/news/press-releases/2021/10/ftc-puts-hundreds-businesses-notice-about-fake-reviews-other-misleading-endorsements.

[13]See https://www.ftc.gov/news-events/news/press-releases/2023/06/federal-trade-commission-announces-proposed-rule-banning-fake-reviews-testimonials. The FTC has also released separate business guidance for both marketers and platforms on online reviews, and is seeking public comment on a potential revision of its pre-existing Endorsement Guides. See https://www.ftc.gov/business-guidance/resources/soliciting-paying-online-reviews-guide-marketers and https://www.ftc.gov/business-guidance/resources/featuring-online-customer-reviews-guide-platforms for, and https://www.ftc.gov/news-events/news/press-releases/2022/05/ftc-proposes-strengthen-advertising-guidelines-against-fake-manipulated-reviews for the request for public comment.

penalties.

The FTC's proposed rule would ban businesses from writing or selling reviews by fictional individuals or those who lack experience with the product or service, and knowingly disseminating fake or misrepresented testimonials. The rule prohibits review hijacking, which is repurposing a review for a different product, as well as offering incentives for positive or negative reviews. Insiders such as officers and managers must disclose their relationships when writing reviews or providing testimonials. Businesses cannot create biased review websites that favor their own products or services. The rule also prohibits preventing or removing negative reviews through unjustified legal threats or intimidation. Finally, the rule prohibits businesses from selling false indicators of social media influence, like fake followers or views.

3 Data

3.1 Sample Construction

My universe of data is the BBB's database of businesses as of February 28, 2020, which included more than 4.8 million unique businesses. In order to isolate businesses with any recent activity on the BBB's platform, I define the sampling frame to include only US located businesses with a BBB letter grade and at least one review or complaint within three years.

I also 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. I thus restrict the sample to businesses with 1,000 employees or less and with fewer than 6 listed

locations. The resulting sampling frame has 628,478 businesses.[14]

I could not match all of these businesses to listings on review platforms due to financial constraints for the Google APIs described in the next section. I thus examine a random sample of businesses developed through a stratified sampling design, oversampling 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.[15] I sample all businesses with at least 10 or more reviews or at least 10 or more complaints, which I use as a proxy for significant activity on the BBB's website.

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 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 high grade and 50,000 businesses with a low grade. Because the sampling frame has many more high grade businesses than low grade businesses, 37.8% of low grade businesses are sampled, compared to 10.8% of high grade businesses.

[14] 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.

[15] 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.

3.2 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 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.[16] 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 three platforms with the largest number of businesses with review ratings: Yelp, Facebook, and HomeAdvisor.[17] I used the Google Places API to provide Google review ratings using the same search string, as the Google Search API does not provide data on reviews on Google's own platform.

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

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

[17] 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.

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 and found a high degree of accuracy of these matches; 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.[18] Finally, I matched data on Yelp listings to data on individual reviews (both published on the website and hidden as they are not "recommended" by Yelp) provided to me by Yelp using the business's website link.

3.3 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.[19] 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.[20] 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

[18] 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.

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

[20] 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.

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

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 7% have an F grade.[21] 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.

4 Ratings and Signals of Quality

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

Figure 1 displays the distribution of average business ratings across platforms. For Google, Facebook, and HomeAdvisor, most businesses have average ratings above 4 stars, with 59% of businesses on Google, 79% on Facebook, and 96% on HomeAdvisor above 4 stars. Only 4% of businesses on Google, 2% on Facebook have an average below 2 stars;

[21] 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.

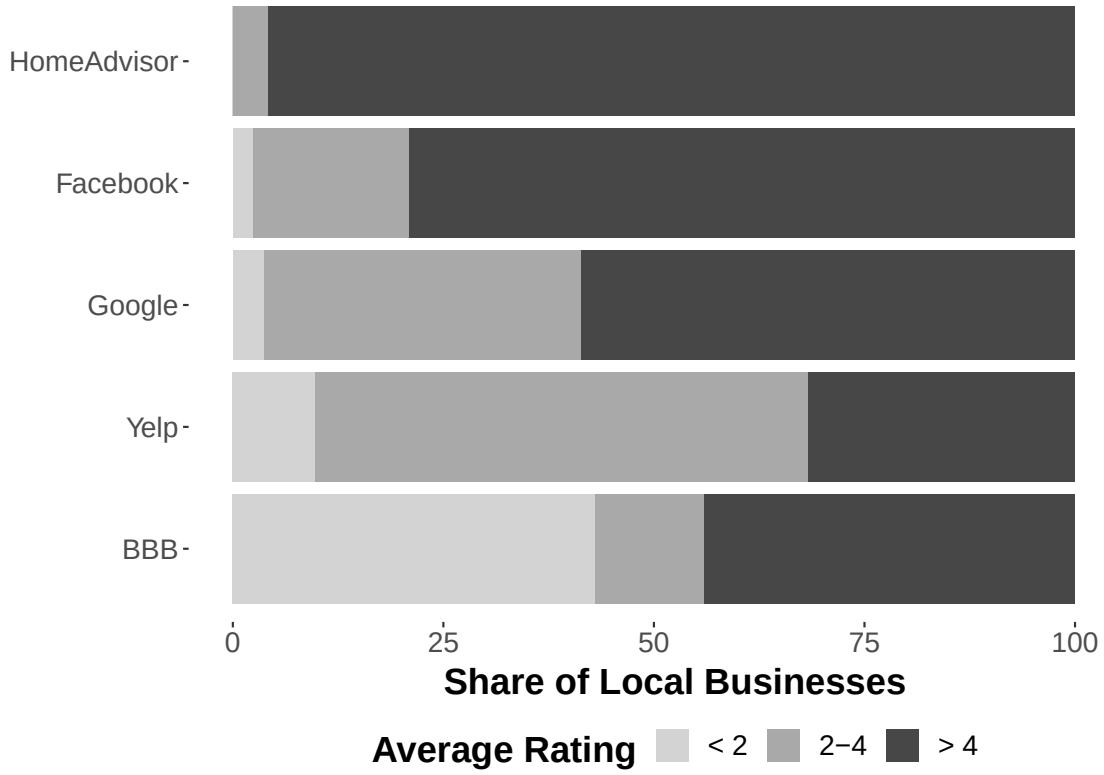


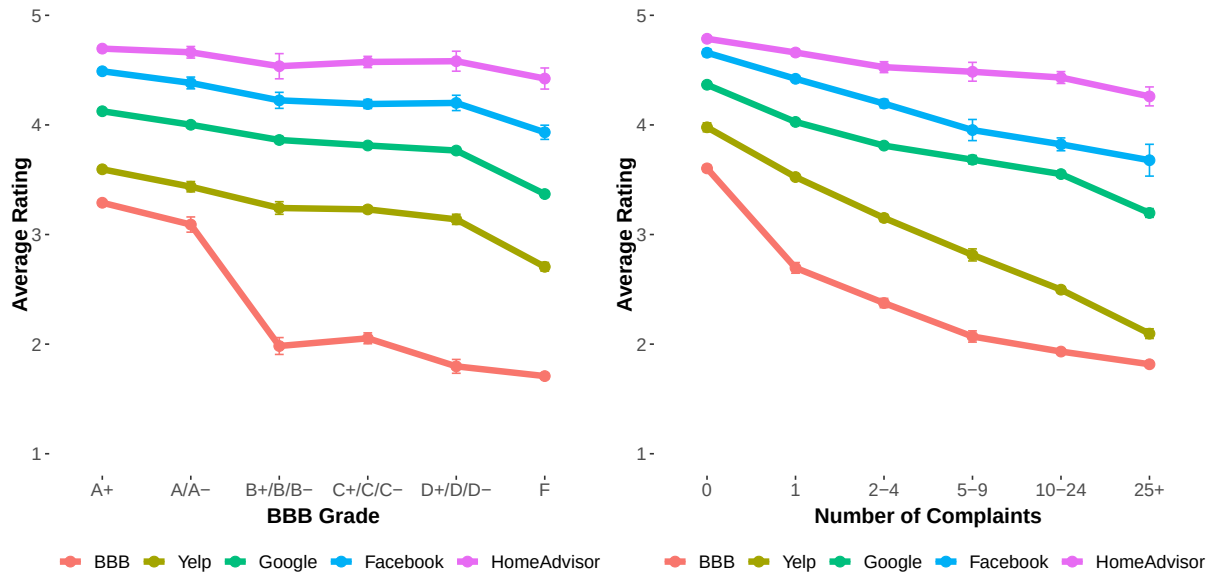
Figure 1 Distribution of Average Business Ratings Across Platforms

Note: All observations weighted using the sampling weights.

no HomeAdvisor businessses are below 2 stars. Like all the children of Garrison Keillor's fictional Lake Wobegon, almost all businesses on Google, Facebook, and HomeAdvisor are above average. Almost none are below average.

In contrast, the BBB ratings are bimodal, with most businesses having either a rating above 4 stars or below 2 stars, and Yelp ratings are much more uniform across the rating distribution. For Yelp, 10% of businesses are below 2 stars and 32% are above 4 stars; for the BBB, 43% are below 2 stars and 44% are above 4 stars. I show similar patterns after controlling for local business effects in Appendix C.1.

Next, I examine how review ratings vary with two signals of quality. First, the BBB assigns grades from A+ to F with plus and minus grades for all letter grades except F; these



(a) BBB Letter Grade

(b) Number of Complaints

Figure 2 Average Rating by Platform and Category

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

grades do not depend upon review ratings.[22] 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 decline of 1.6 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, 0.9 stars for Yelp, 0.8 for Google, 0.6 for Facebook, and 0.3 for HomeAdvisor.

Second, I examine how review ratings vary by the number of BBB complaints for the

[22]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/.

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 2b.

The average rating declines with more complaints on all platforms, but the magnitude of this decline is much larger for the BBB and Yelp compared to the other platforms. A business with 25 or more complaints has, on average, a 1.8 star lower rating than a zero complaint business on the BBB and 1.9 star lower rating on Yelp. The decline in rating going from 0 complaints to 25 or more complaints is significantly lower on the other platforms, at, on average, 1.2 for Google, 1.0 for Facebook, and 0.5 for HomeAdvisor. In Appendix C.1, I show that controlling for local business effects does not alter these differences across platforms.

5 Estimating Quality

In this section, I structurally estimate the quality of a business using a finite mixture model, and then examine how ratings of different platforms vary by quality tier.

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, as all of the review platform ask consumers to assign a business a score from one to five (five tiers). The BBB grade is a set of thirteen tiers in the letter grades from A to F with plus and minus gradations. Second, policymakers need to identify businesses with likely consumer protection problems, which might comprise the lowest quality tier. The FTC has previously used finite mixture models to identify businesses that were likely to have committed fraud (Balan et al., 2015).

Under the finite mixture model, the likelihood for business i is:

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

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 distribution of signal k is f_{jk}.

Crucially, 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).[23]

[23]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).

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.

5.1 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).[24] 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. Levine et al. (2011) prove EM-like descent properties for the algorithm, which iterates until convergence. To identify types, I assign each business the type with the highest 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 4: 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

[24] 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.

number of BBB complaints from 2010 to 2016, a seven year period prior to the complaint measure examined in Section 4. 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 III, 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

Table III 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.

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 using the share of businesses deemed as high risk by the BBB, which was not explicitly used in the model as a signal. High risk businesses include pyramid schemes and work at home companies.[25] 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%. The share of high risk businesses increases when quality falls, as one would expect.

Figure 3a 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 to low quality business; the average rating falls from 3.9 for high quality businesses to 3.4 for

[25]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.

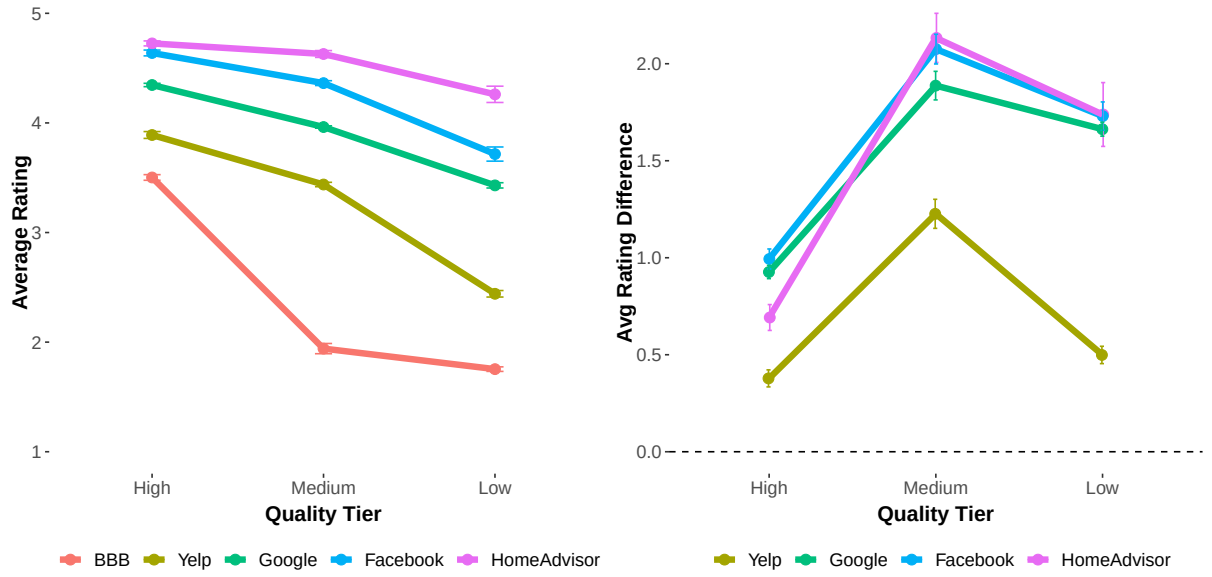
medium quality businesses to 2.4 for low quality businesses. Ratings on Google, Facebook, and HomeAdvisor all decline much less when business quality falls; Google and Facebook's ratings falls by about 0.9 stars on average, and HomeAdvisor's ratings by 0.5 stars, when going from a high quality to low 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 3b depicts the estimates from panel regressions controlling for business fixed effects; as in Section 4, these results are relative to the BBB's rating. In Appendix C.2, 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 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.

5.2 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



(a) Average Rating

(b) Rating Relative to BBB

Figure 3 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.

results to consumers, which may depend upon relative rankings. 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 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 IV 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

Table IV 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.

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. Taken together, the Google, Facebook, and HomeAdvisor platforms have a much more compressed distribution of ratings than the BBB and Yelp, but provide similar relative rankings.

5.3 Alternative Mixture Models

I used a mixture model with 3 types above. With four types, the medium quality tier is separated into two types. With five types, 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 types. Since much of the consumer protection interest is in identifying bad businesses, which in the model correspond to the low quality type, I use three tiers for my main analysis.

I also examine excluding the review ratings as signals in Appendix C.3. 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 as low quality if they have many complaints and low platform ratings.

6 Explanations

There are several potential explanations behind the patterns shown in the previous two sections. First, the types of users reviewing may vary across platforms. For example, Google's reviews are embedded in Google Maps, which most Americans use for directions on their smartphone; Google Maps (or the Android platform more generally) may prompt users to provide reviews. Reviewing on specialist websites such as the BBB or Yelp would require consumers to go to a website or app in order to review. Second, Facebook is primarly a social media platform as opposed to a review site. Reciprocity norms might mean that consumers give positive reviews to businesses in their social network.

Third, the competition concerns described in Section 2 could lead to lower quality reviews from gatekeeper platforms. In Section D, I develop a model of platform steering and quality in which a gatekeeper and independent platform compete for consumers along a Hotelling line, and firms have to invest in quality. Steering that makes the gatekeeper the only considered choice for many consumers lowers the quality of both the gatekeeper and independent platform. Steering that increases the perceived utility of the gatekeeper – such

as preferential placement and visuals – decreases the quality of the gatekeeper relative to the independent. On the other hand, if the gatekeeper receives more profit per user than the independent platform, it would invest in higher quality than the independent.

I indeed find higher ratings for low quality businesses for the gatekeeper platforms Google and Facebook compared to the BBB and Yelp, which are independent platforms. However, I find similar review rating patterns to Google and Facebook for HomeAdvisor, which is independent.[26]

While I cannot fully examine all of the potential explanations behind differences in reviews across plaforms, I do evaluate two explanations for differences across platforms in this section. One explanation is differences in the prevalance of fake reviews across platforms; here, I have proxies for whether a review is fake for the BBB and Yelp and can examine how star ratings vary between reviews more likely to be real and more likely to be fake. Second, platforms can decide on how much effort consumers have to put to write a review, such as the amount of text required. For Google and Yelp, I examine how ratings distributions vary by the amount of text that consumers write.

6.1 Fake Reviews

I proxy for fake reviews using data from the review filtering algorithms of the BBB and 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 and not included in the average rating.

[26]The FTC recently settled allegations that HomeAdvisor misled businesses on the quality of leads they received from consumers seeking home improvement professionals for up to $7.2 million. These allegations might reflect poor quality generally for HomeAdvisor, or incentives to inflate reviews of such businesses to increase the number of leads from consumers. See https://www.ftc.gov/legal-library/browse/cases-proceedings/1923106-homeadvisor-matter. for details on the case.

Luca and Zervas (2016) provide evidence that these hidden reviews are a good proxy for fake reviews. In addition, I use data on the BBB's own proprietary filtering algorithm's score for each review, including both published and unpublished reviews, meant to predict the probability that a review is fake.[27] I use this score 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.

I examine two channels on how fake reviews could affect how ratings reflect business quality: the difference in rating between fake reviews and real reviews by type of business, and the share of fake reviews by type of business. Table V 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. These results could mean that all businesses have fake reviews, or that creators of fake reviews are good enough at spoofing real reviews that many legitimate reviews of high quality firms are flagged as fake.

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.

[27] 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 V Share of Likely Fake Reviews by Quality Tier

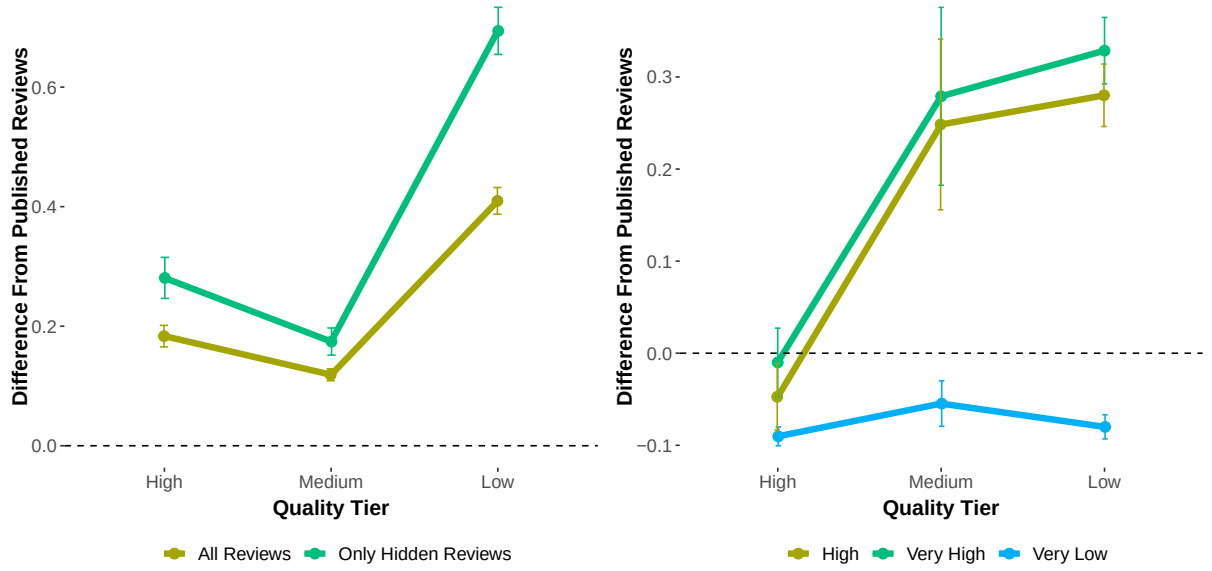
	BBB Very Likely to Be Fake	Yelp Hidden
High	7.9	46.6
	(0.2)	(0.4)
Medium	8.9	33.1
	(0.2)	(0.2)
Low	9.4	46.0
	(0.2)	(0.4)
Observations	72257	25792

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

For Yelp, I examine ratings using all reviews – hidden and published – as well as only hidden reviews. I depict these results in Figure 4a. 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.

Review ratings based on the fake review probability from the BBB filtering algorithm have a similar pattern to the Yelp results; Figure 4b displays the BBB 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 contrast, 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



(a) Yelp Publication Status

(b) BBB Fake Review Probability

Figure 4 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.

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 5. Average ratings based upon likely fake BBB and Yelp reviews are about 1 star and 0.3 stars lower than Google ratings, whereas average ratings for published reviews are 1.7 and 1 stars lower than Google ratings.

6.2 Review Effort

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 effort that users on those platforms spend to write a review. I focus on the length of reviews,

which is partially a decision variable of the platform, as a measure of effort.

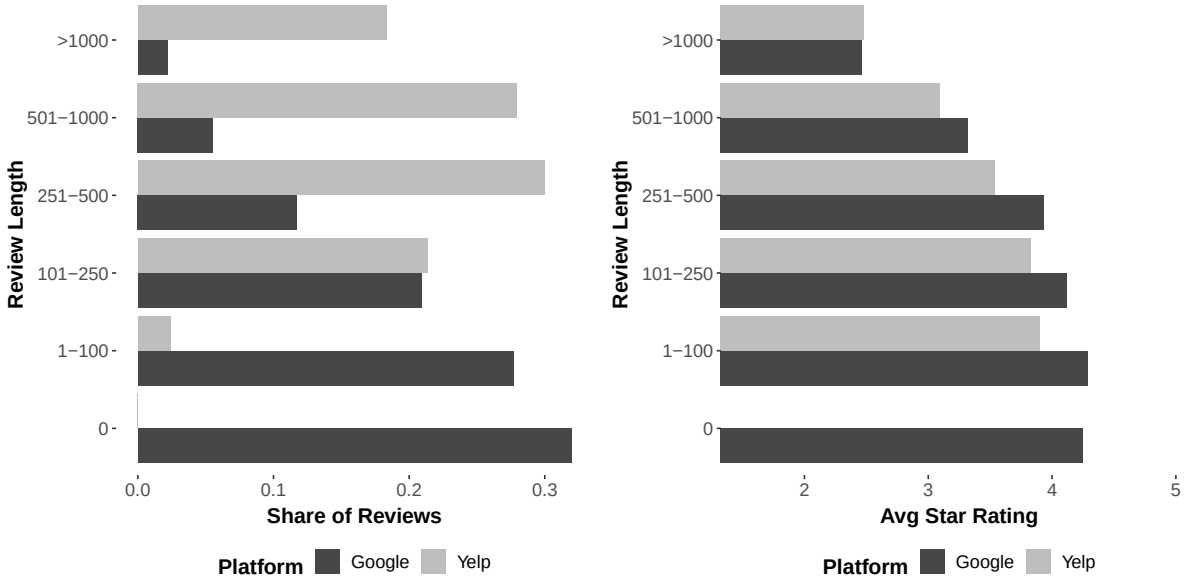
Platforms vary substantially in their policies on review length, and have changed these policies over time. Yelp and the BBB do not allow ratings without any review text, while Google does allow such "no-text" ratings. Facebook used to allow no-text ratings, but now imposes a 25 character limit. With 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 how review length might affect ratings by comparing Google and Yelp reviews. Yelp provided me data on all reviews for the businesses in the sample. For Google, I scraped data on reviews for businesses in the sample several months after the initial data collection. In total, I was able to scrape reviews for about 87% of businesses with matched Google listings in the original dataset. About half of the unmatched businesses had closed since the original data collection, while for half I could not scrape all of the reviews.[28]

Reviews on Yelp are, on average, much longer than those on Google. Figure 5a displays the share of reviews on both platform by review length category. For Google, 32% of reviews have no text. In addition, 28% of reviews on Google have between 1 and 100 characters, compared to 2% of reviews on Yelp with less than 100 characters. On the other hand, 6% of reviews on Google have 501 to 1,000 characters, and 2% have more than 1,000 characters, compared to 28% of Yelp reviews with 501 to 1,000 characters and 18% with more than

[28]Google only displays up to 930 reviews for a business on its website for a given review ordering. I thus scraped reviews ordering reviews by highest rating and separately by lowest rating, and exclude businesses for which I could not scrape all of the reviews. I then remove reviews posted after the initial data collection, using the recorded date of review posting.



(a) Share of Reviews

(b) Average Rating

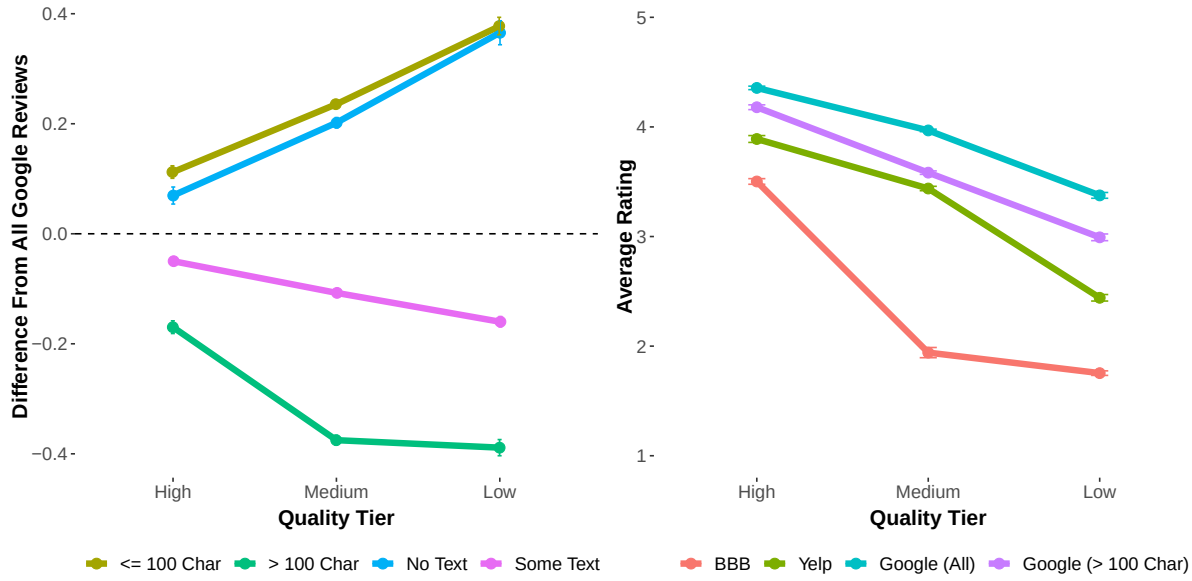
Figure 5 Share of Reviews and Average Rating by Review Length and Platform

Note: Estimates use reviews for all businesses in the sample weighted using the sampling weights. For Google, reviews are limited to businesses for which I could scrape all the reviews, as described in the text.

1,000 characters.

For both Google and Yelp, longer reviews have, on average, lower ratings. Figure 5b displays the average star rating on both platforms by review length category. No-text reviews on Google have an average of 4.3 stars; reviews with 1 to 100 characters have an average of 4.3 stars on Google and 3.9 stars on Yelp. In contrast, reviews with 501 to 1,000 characters have 3.3 stars on Google and 3.1 stars on Yelp, and reviews with more than 1,000 characters have an average of 2.5 stars on Google and Yelp.

In Figure 6a, I compare how the average rating would change for Google by only including certain reviews based on review length. I estimate panel regressions that control for individual business fixed effects, comparing ratings based on alternative review lengths for Google to all ratings for different quality tiers. Short reviews have higher ratings, and



(a) Rating Relative to All Google Reviews

(b) Average Rating

Figure 6 Rating by Platform and Quality Tier By Character Length of Reviews

Note: Estimates clustered at the individual business level. Estimates use all businesses in the sample weighted using the sampling weights. For Google, reviews are limited to businesses for which I could scrape all the reviews, as described in the text.

long reviews have lower ratings, compared to all reviews across quality tiers. For example, restricting reviews to only those above 100 characters decreases the average rating by 0.2 stars for high quality businesses and 0.4 stars for medium and low quality businesses.

In Figure 6b, I depict the average rating by quality tier for all Google reviews and Google reviews above 100 characters and compare to the BBB and Yelp. Removing short Google reviews substantially reduces the gap between Yelp and Google, although the gap for low quality businesses remains about double the gap for high quality reviews.

7 Conclusion

In this article, I have compared five different review platforms on how their ratings of local businesses correspond to business quality. To measure the quality of a business reviewed on a platform, I estimated a finite mixture model incorporating several signals of consumer protection problems. While Google, Facebook, and HomeAdvisor have higher ratings on average than the BBB and Yelp, this gap in average rating is the highest for low quality businesses. However, low quality businesses almost always have a lower rating than high quality businesses on all the platforms.

I also found evidence that fake reviews can explain some of these differences. 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 but not high quality businesses. In addition, platforms vary in their policies on the length of reviews. Google has many more reviews than Yelp, but about half of Google reviews have 100 characters or less. Removing reviews with little text reduces ratings for both low quality and high quality businesses, and so is less likely to account for the disparity between platforms for low quality businesses.

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 regulators, this research has shown how online reviews vary across competing platforms, and in particular between dominant platforms and their independent rivals. In addition, this work has shown that policing fake reviews is valuable, as fake reviews disproportionately boost ratings of low quality business with consumer protection problems. 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.[29]

[29]For example, Raval (2020) documents substantial selection in consumers who choose to complain, with

8 Funding and Competing Interests

I have no competing interests.

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.

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

Fang, Limin, "The effects of online review platforms on restaurant revenue, consumer learning, and welfare," *Management Science*, 2022, *68* (11), 8116–8143.

Fradkin, Andrey and David Holtz, "Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb," *Marketing Science*, forthcoming.

_ , Elena Grewal, and David Holtz, "Reciprocity and Unveiling in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb," *Marketing Science*, 2021, *40* (6).

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.

victims in Black and Hispanic areas much less likely to complain about fraud.

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.

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.

Nadler, Jerrold and David N. Cicilline, "Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations," Technical Report, US House of Representatives Subcommittee on Antitrust, Commercial, and Administrative Law of the Committee of the Judiciary 2020.

Nosko, Chris and Steven Tadelis, "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment," Technical Report, National Bureau of Economic Research 2015.

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 Truth—Experimental 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.

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.

Wang, Chengsi and Julian Wright, "Search platforms: showrooming and price parity clauses," *The RAND Journal of Economics*, 2020, *51* (1), 32–58.

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.

Zervas, Georgios, Davide Proserpio, and John W Byers, "A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average," *Marketing Letters*, 2021, *32* (1), 1–16.

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

Platform	Categories Matching: 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.[30]

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).

B Levine et al. (2011) Algorithm

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

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

[30]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).

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).

C Empirical Appendix

C.1 Ratings and Quality Signals

In this section, I provide tables of the average rating and distribution of ratings by platform, how these results change after adding local business controls, as well as how the correlation between

Table A-2 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, 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.[31]

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

$$Y_{ip} = \gamma_p + \delta_i + \epsilon_{ip}, \tag{2}$$

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.

[31] 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 A-2 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

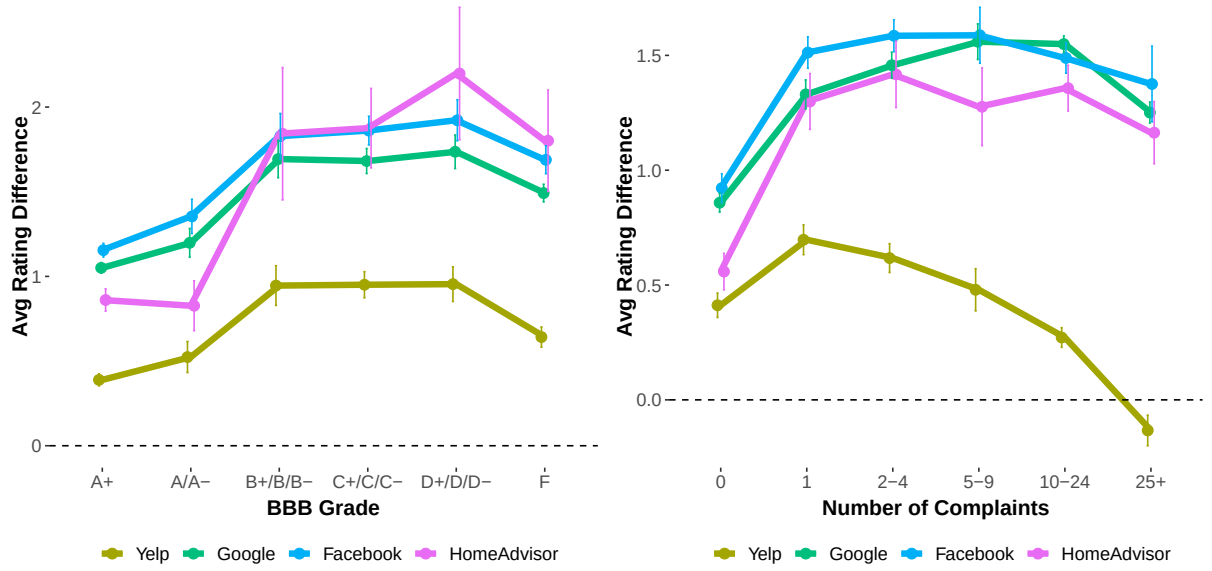
Note: Estimates clustered at the individual business level and include all businesses in the sample weighted using the sampling weights.

Table A-3 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 A-3 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.



(a) Rating Relative to BBB

(b) Rating Relative to BBB

Figure 7 Average Rating Relative to BBB by Platform and Category

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

C.1.1 BBB Letter Grade

Next, I examine how review ratings vary by the BBB letter grade for the business. 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:

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

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 7a 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 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.

C.1.2 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 2b.

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:

$$Y_{ip} = \alpha_{c(i)p} + \delta_i + \epsilon_{ip}, \tag{4}$$

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 7b 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.

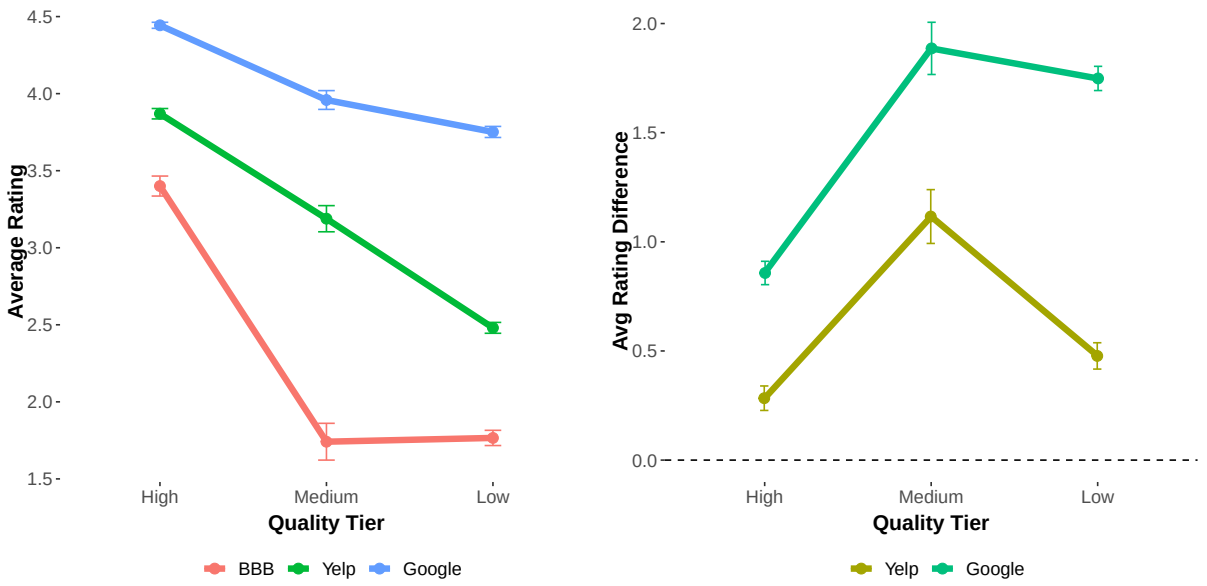
A business with zero complaints has a 0.41 star higher rating on Yelp than the BBB, a 0.86 star higher rating on Google, a 0.92 higher star rating on Facebook, and a 0.56 higher star rating on HomeAdvisor. For businesses with 25 or more complaints, Yelp ratings are 0.1 stars lower than BBB ratings. In 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 stars higher.

C.2 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.

C.3 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-4 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



(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.

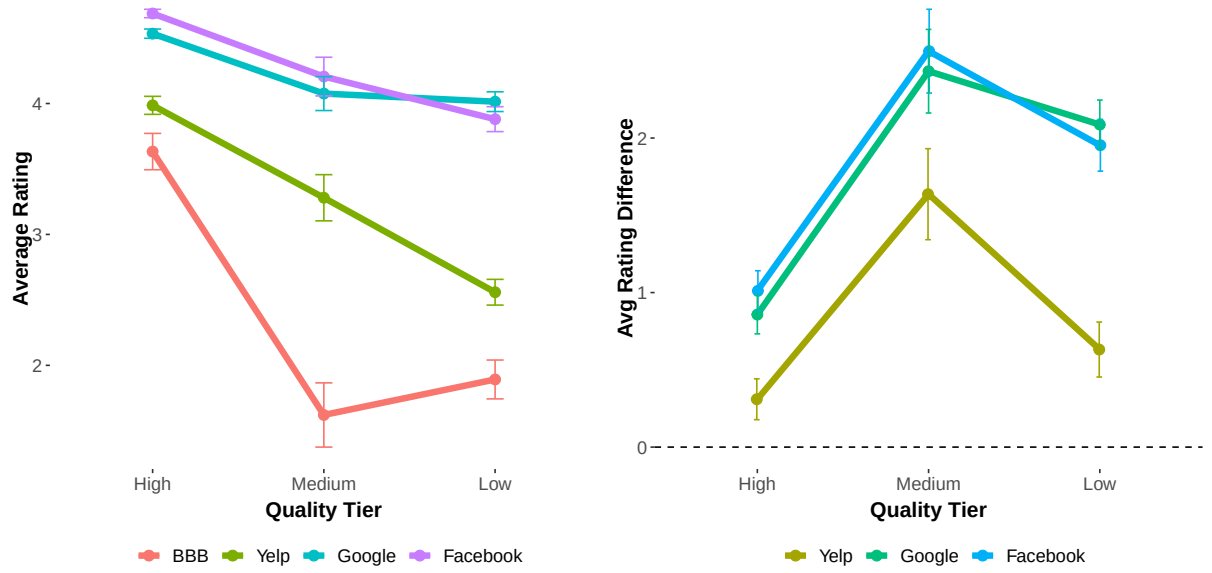
controlling for business fixed effects; as in Section 4, these results are relative to the BBB's rating. Estimates using this quality tier are similar to those reported in the text using all 10 signals, except the difference between Yelp and Google for low quality businesses is smaller.

C.4 Review Length

In this section, I examine questions of review length using two auxiliary datasets that are large corpuses of reviews. 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.[32] 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 11a displays the share of reviews on both platform by review length 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,

[32]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.



(a) Average Rating

(b) Rating Relative to BBB

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.

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 11b 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.

Table A-4 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.

D Model

In this section, I build a simple model of platform steering with both a gatekeeper and independent platform, and show that platform steering can reduce the quality of both platforms.

Two platforms compete for customers located uniformly on a Hotelling line of length 1, with the two platforms located on either end of the line. The utility that customers receive from each platform is:

$$u(\alpha + \theta s_1) - tx \tag{5}$$

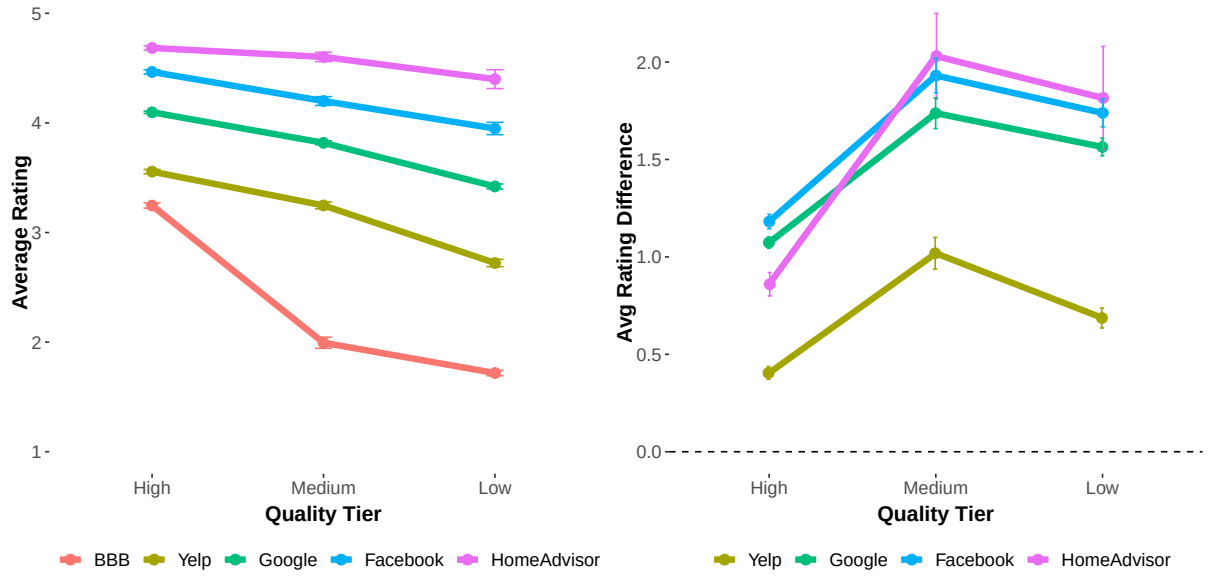
$$u(\theta s_2) - t(1 - x) \tag{6}$$

Here, s_1 and s_2 are the qualities chosen by platforms 1 and 2 respectively, with θ the consumer valuation of quality. The consumer's utility function for the good is u. Consumers also face transport costs t based on their distance from the platform x, which provides horizontal differentiation between the two platforms.

Platform 1, the gatekeeper platform, has two mechanisms for steering, which depend upon whether consumers observe the independent platform. First, α is a steering parameter which enters consumers' utility additively, so consumers see the gatekeeper's product as superior to the independent's product at equal quality. One example of such steering would be placement on a search page. If the gatekeeper places its own review results above rival platforms, and consumers give more weight to higher ranked results, the gatekeeper's conduct would have inflated consumers' valuation of the gatekeeper's product. Another example would be a user interface that recommends the gatekeeper's product explicitly through a badge (e.g., Amazon's Choice) or implicitly through a special text box with pictures and additional content.

Second, a β fraction of consumers do not search and stay on the gatekeeper platform's website, and so automatically choose platform 1 (Wang and Wright, 2020). For example, consumers using Google Maps to navigate and search for businesses will only see Google reviews on the app, and have to actively search by switching to another app to see reviews from other platforms. In other words, only $1 - \beta$ of the market is contestable.

In the game each platform first chooses its quality s, and then consumers decide which platform to use. I start with the consumer's problem given platforms' choice of quality for the $1 - \beta$ fraction of consumers making an active choice between platforms.



(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.

A consumer is indifferent between platforms if their utility from each platform is the same, and so an indifferent consumer has distance from platform 1 x^* as follows:

$$u(\alpha + \theta s_1) - tx^* = u(\theta s_2) - t(1 - x^*) \tag{7}$$

$$x^* = \frac{1}{2} + \frac{u(\alpha + \theta s_1) - u(\theta s_2)}{2t} \tag{8}$$

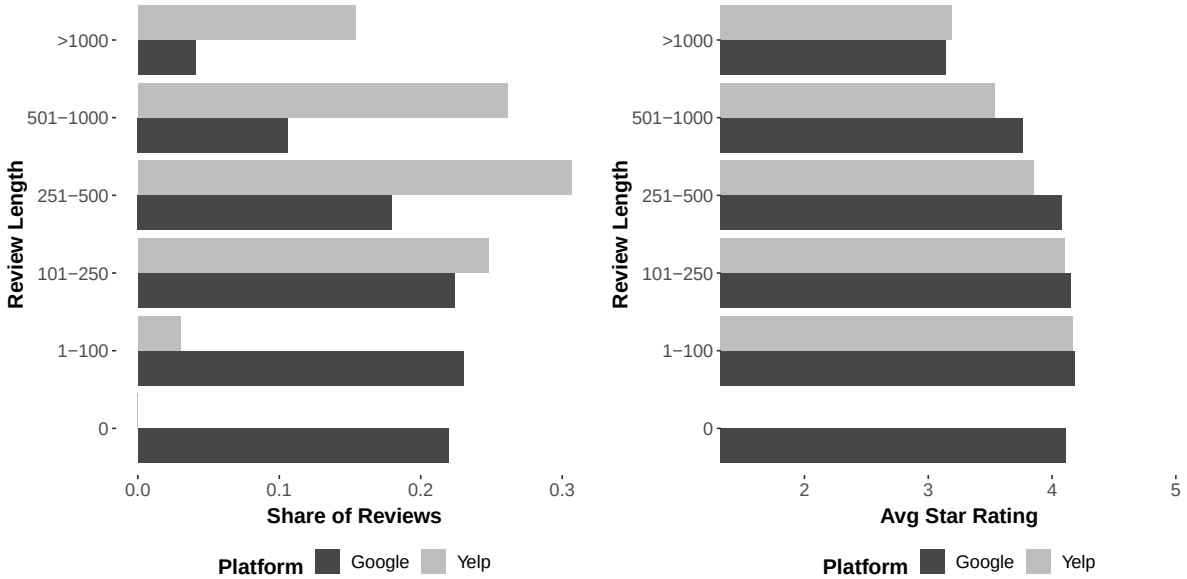
The demand for platform 1 is then $\beta + (1 - \beta)x^*$ and for platform 2 $(1 - \beta)(1 - x^*)$.

In stage 1, each platform then chooses a level of quality. I assume that platforms earn p_1 and p_2, respectively, per user from advertising and pay a cost of quality $c(s)$.[33] In that case, each platform maximizes its profits as follows:

$$\max_{s_1} p_1[\beta + (1 - \beta)x^*] - c(s_1) \tag{9}$$

$$\max_{s_2} p_2(1 - \beta)(1 - x^*) - c(s_2) \tag{10}$$

[33]While revenue here is modeled per user, platforms could obtain revenue from businesses advertising to reach those users.



(a) Share of Reviews

(b) Average Rating

Figure 11 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.

The optimal choice of s_1 and s_2 then satisfy:

$$\frac{p_1(1-\beta)\theta}{2t}u'(\alpha+\theta s_1) = c'(s_1) \tag{11}$$

$$\frac{p_2(1-\beta)\theta}{2t}u'(\theta s_2) = c'(s_2) \tag{12}$$

The degree of steering α does not affect the independent platform's decision on quality. Using the implicit function theorem, the effect of the degree of steering α on gatekeeper platform 1's quality s_1 is:

$$\frac{ds_1}{d\alpha} = \frac{\theta(1-\beta)p_1u''}{(2tc'' - p_1(1-\beta)\theta^2u'')} \tag{13}$$

The effect of steering on platform 1's quality is negative (i.e. $\frac{ds_1}{d\alpha} < 0$) if the cost function for quality is weakly convex ($c''(s) \geq 0$) and the utility function is concave ($u''() < 0$). With these same conditions, an increase in β (i.e. a decrease in the contestable market) decreases both platforms' quality, and an increase in a platform's price increases its quality.[34]

If the cost function is linear, we have the particular simple result that:

$$\frac{ds_1}{d\alpha} = -\frac{1}{\theta} < 0 \tag{14}$$

[34] $\frac{ds_1}{d\beta} = \frac{-\theta p_1 u'}{(2tc''-p_1(1-\beta)\theta^2u'')} < 0$ and $\frac{ds_1}{dp_1} = \frac{\theta(1-\beta)u'}{(2tc''-p_1(1-\beta)\theta^2u'')} > 0$.

Thus, increasing the steering parameter α lowers the gatekeeper platform's choice of quality provided that the consumer's utility function is concave.

I now assume that the utility function is the natural log and the cost of quality is linear, so $c(s) = is$, to obtain simple closed form solutions for the quality of each platform:

$$s_1 = \frac{p_1(1-\beta)}{2it} - \frac{\alpha}{\theta} \tag{15}$$

$$s_2 = \frac{p_2(1-\beta)}{2it} \tag{16}$$

$$s_1 - s_2 = \frac{(p_1 - p_2)(1-\beta)}{2it} - \frac{\alpha}{\theta} \tag{17}$$

Here, the gap between the quality of the gatekeeper platform and alternative platform is increasing in the degree of steering α, and decreasing in consumer's valuation of quality θ. Increasing β, so the contestable market is smaller, decreases the quality of both platforms.

Both platforms' quality increases in the price they receive per consumer p, and decreases in the investment cost term i and the degree of horizontal differentiation t. The gatekeeper might receive a higher value for consumer, perhaps because it can match consumer behavior to other information on the user, or because using the platform leads to more platform engagement generally. In that scenario, the gatekeeper platform could choose higher quality than the independent platform. Thus, whether the gatekeeper has higher or lower quality than the incumbent is an empirical question.

How is welfare affected by steering? If we judge welfare based on consumers without any steering effects, setting α and β to zero would unambiguously increase consumer utility, as quality would improve on both platforms.

However, effects on total welfare will depend upon whether the improvement in consumer utility outweighs the increased cost of additional quality. Take, for example, the case of log utility and linear cost. The increase in quality for the gatekeeper platform after setting α to zero will be $\frac{\alpha}{\theta}$ at cost $\frac{i\alpha}{\theta}$, for any price of advertising or level of horizontal differentiation. The increase in utility of consumers from this improvement in quality does depend on these factors; it will decrease as prices rise or platforms are less horizontally differentiated, as then overall quality is higher and so the benefit to consumers of increased quality is lower.

In addition, the simple model above treats the steering parameters α and β as exogenous. In reality, the gatekeeper may be able to increase the degree of steering at some cost. For example, adding more advertising links could increase the distance between the gatekeeper's search results and rivals, and improving other parts of the gatekeeper's platform (e.g. Google Maps) could increase the share of consumers who remain on the app and do not search.