Chasing Stars: Firms' Strategic Responses to Online Consumer Ratings[*]

Megan Hunter[†]

June 22, 2020

Working Draft

Latest Version Available Here

Abstract

In this paper, I show that a common way that platforms display firms' quality ratings incentivizes firms to strategically take costly short-run actions that improve their ratings. Most review platforms display star ratings of goods and services rounded to a half star, rather than display the exact average rating. Since the true average rating is not shown, firms have an incentive to remain just above the rounding threshold in order to have a higher displayed rating. However, once a firm's rating passes the rounding threshold, the incentive to improve the ratings drops as their rating moves farther from the threshold. I study this phenomena in the context of auto repair. Consumers face significant uncertainty in the auto-repair market, which makes it a prime context in which to study reviews. I first show that consumers respond to ratings in this market. Firms with higher-displayed ratings have higher revenues and more customers than lower-rated firms. To identify the effects of ratings versus quality, I utilize a regression-discontinuity strategy. Since ratings have a significant effect on demand, firms have an incentive to pay attention to their ratings. Consistent with

[*] I would like to acknowledge my deepest thanks to my advisor, Wes Hartmann for his dedication and feedback throughout this project and my entire PhD. I would also like to thank my committee members, Lanier Benkard and James Lattin, for their time and wise help. Additionally, I thank Stanford GSB Marketing faculty and students, Sam Antill, Brad Larsen, Sridhar Moorthy, Sridhar Narayanan, and Peter Reiss for helpful comments.

[†] Boston College Carroll School of Management, megan.hunter@bc.edu

the firms working to strategically push their ratings past rounding thresholds and keep them above, I find that there is an excessive amount of bunching around ratings thresholds. The firms' actions toward improving their ratings are typically unobserved, but due to my novel data and the discontinuity of displayed ratings, I can model and infer firm behavior. Specifically, I provide evidence that firms change the services they provide and exert extra effort when they are close to rounding thresholds. Finally, I provide a theoretical framework in order to quantify the actions and provide optimal policies for firm actions depending on their rating and number of reviews.

1 Introduction

Consumer reviews have become a standard feature in most industries; consumers can find reviews for almost any product or service. Many consumers read reviews; up to 90% of consumers read online reviews before making a purchase.[1] There is a large marketing literature on how consumers use reviews and ratings when searching and making purchase decisions. However, the literature on the supply-side response to reviews is much more sparse. More broadly, when considering firms' actions, the marketing literature has traditionally been split between studying short-run changes, such as pricing or digital advertising, and long-run actions, such as which products to produce or inherent quality levels. There are also short-run actions that firms can take that are usually unobservable. I exploit the nature of how ratings are displayed online to identify a model of these unobserved actions.

This paper seeks to answer the following questions. Do firms engage in previously unobserved short-run strategic activities in order to influence their ratings? How do these actions vary quality perceptions or change the distribution of quality available? Does a firm's incentives change depending on: (1) how ratings are displayed to consumers and (2) the firm's rating state, which is a function of the firm's mean rating and number of reviews? How does the way that ratings are displayed affect consumer welfare?

To answer my research questions, I first document that consumers respond to reviews in my setting by showing that higher ratings are associated with increased demand, both in revenue and number of customers. I find that ratings matter more in competitive markets and for new customers, which is consistent with the information consumers gain from reviews. I then provide evidence that firms are responding deliberately to their online reviews, particularly when their ratings are close to the ratings-rounding thresholds. Firms have an incentive to move their average rating just past rounding thresholds in order to be displayed as a higher star rating to consumers. Finally, I create a model of a firm's strategy in order

[1] https://www.forbes.com/sites/ryanerskine/2017/09/19/20-online-reputation-statistics-that-every-business-owner-needs-to-know#542c5c52cc5c

to simulate how firms would change their behavior if ratings were displayed differently to consumers.

The way in which ratings are commonly displayed by websites and platforms not only impacts the behavior of firms, but also makes it possible for me to identify the actions that firms take to improve their ratings. Most review platforms display star ratings of goods and services rounded to a half star, rather than display the exact average rating. In particular, Yelp, Amazon, and TripAdvisor, some of the largest review platforms,[2] [3] display ratings on a star scale of one to five and round to the nearest half star. The displayed ratings are then one of the following: 1 star, 1.5 stars, 2 stars, 2.5 stars, 3 stars, 3.5 stars, 4 stars, 4.5 stars or 5 stars. These ratings are thus discontinuous with respect to the average rating. This discontinuity leads to a particular incentive structure for firms, as their displayed rating is much more important than their average rating. Firms know that consumers use ratings when choosing what product to buy or from which website to purchase. Therefore, firms are likely to respond accordingly in order to obtain strong positive reviews from their consumers in hopes that this will increase their future revenue stream. Firms can lower their product's price, better their customer service, change their advertising strategy, or try to avoid customers or services that they expect will result in lower ratings. An equilibrium in these effects exists, and the rounding of displayed reviews provides a unique lens to understand firms' incentives. Firms' responses are likely to vary depending on the current ratings, both as a function of their average rating and the number of reviews.

These phenomena and ratings responses can be found across a variety of industries. A simplified example of this is an Uber driver. Uber's ratings are continuous, they do not have these displayed stars. However, it has been noted that there is a pass/fail cutoff; Uber

[2] https://magazine.startus.cc/the-review-platforms-your-business-needs-in-2018/

[3] http://www.yourtechstory.com/2019/12/10/unleash-the-passionate-traveller-inside-you-with-tripadvisor/

drivers can be removed from the Uber platform[4] if their ratings drop below a 4.6.[5] If a driver's rating is just above this cutoff, say a 4.61, the driver might change her behavior. For example, if an Uber driver receives a ride request from a passenger with low ratings, the driver might assume that they are also the type of passenger to give low ratings in return, and the driver may turn down the ride.[6] Alternatively, she may put in extra effort to the ride, taking music requests or providing water bottles.[7] It is likely that the driver will be more likely to take on these strategies when her own rating is getting close to hitting the 4.6 threshold as a new low rating could drop her below the threshold and get her kicked off.

My particular context is the auto-repair industry. The auto-repair market is quite large; in 2017, the industry had annual sales of $63 billion in the United States.[8] The auto-repair industry is a well suited market to study these research questions because auto repair is a large purchase item for consumers and a credence good that the consumer likely does not know much about. Therefore, consumers are likely to turn toward additional pieces of information and firms need to build their reputation.

The first part of the paper explores how consumers respond to reviews in my setting. An inherent issue in consumer-ratings analysis is to disentangle the effects of quality and rating, as it is likely that high-quality firms and products are also rated highly. I exploit the discontinuity in displayed rating, as a function of the average rating, for identification of the quality and rating issue as first implemented by Luca 2016. For example, on Yelp, an auto-repair shop with a 4.24 average rating will be displayed as a 4-star shop, but a shop with a 4.26 average rating will be displayed as a 4.5-star shop. These auto-repair shops are likely to be similar in most aspects, including quality, but by chance were rated slightly differently

[4] Officially on Uber's website, "If your rating approaches the minimum for your area, you'll receive notifications and tips for how to improve it. If your average rating continues to fall below the minimum after multiple notifications, your account may be deactivated pursuant to the Community Guidelines. Deactivation is only used as a last resort and your account may be activated if you take certain steps to improve." https://www.uber.com/drive/resources/how-ratings-work/

[5] https://www.businessinsider.com/how-uber-drivers-get-deactivated-2017-7

[6] While Uber drivers cannot cancel too often, the same Business Insider article mentions that a driver needs to keep their cancellation rate below 10% so there is room for cancellation.

[7] I heard of a driver who would give his passengers a $5 bill if they left him a 5-star review.

[8] https://autodub.com/automotive-repair-market-usa/

and thus displayed as different star ratings, through which the consumer infers different levels of quality. By comparing shops that are just below and just above the displayed rounding cutoff, I can assume that the shops are relatively similar and comparable, except for the displayed rating that the consumer sees, and therefore isolate the effect of ratings on demand.

Next, I focus on the firm response to reviews — how, and in what rating states, do firms try to improve their ratings? I provide several pieces of evidence that firms are responding to their reviews when they are close to the rounding thresholds. First, the rate at which reviews are left increases when a firm's average rating is within a certain bandwidth around these rounding thresholds. The rate of review incidence increases even more when a firm's rating is just below the threshold. Second, the average ratings received are higher within this bandwidth. Third, I note an excessive amount of bunching of reviews just above a rounding threshold for a displayed star rating, and a trough just below the rounding threshold, as seen in Figure 1. This excess mass is greater than expected, and indicates that there is something a firm can do in order to increase the number of reviews that they obtain right when their review is being pushed across a rounding threshold. In other words, firms take on some strategic actions to improve their ratings just enough to have a higher displayed rating.

I consider two potential strategies that firms can take: turning away customers or types of services that have previously lead to poor reviews, or exerting extra (costly) effort. Firms are more likely to engage in these strategies when their ratings are close to the rounding thresholds. In particular, since turning away customers prevents poor ratings, this strategy is more likely to be used when a shop's average rating is just above a rounding threshold. Exerting extra effort encourages a positive review, and is thus more likely to be used when a shop's average rating is just below a threshold. Both of these responses are changes that firms can make quickly, allowing them to respond as reviews arrive. The second strategy, of exerting extra effort, cannot be directly observed, and thus I create a structural model in

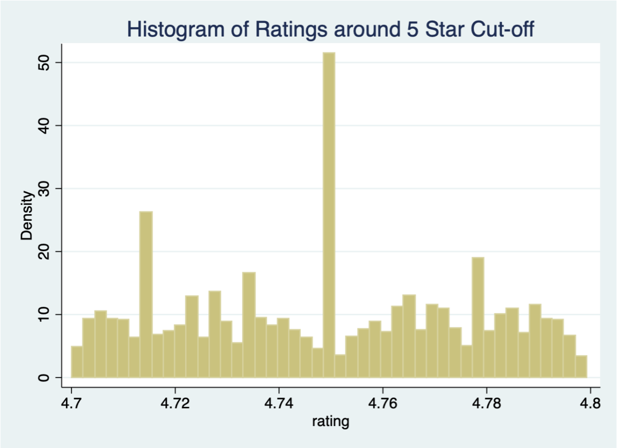


Figure 1: Ratings Distribution: An excess amount of mass can be seen at 4.75. There are additional spikes due to the fact that low numbers of reviews are more common and only certain averages can occur with certain numbers of reviews; however, the excess bunching at the threshold is robust to this.

order to back out this activity.

I first consider how a given repair shop changes the services it performs in response to their ratings. Using a novel feature of my data set, I can match individual ratings to specific invoices and therefore determine ratings for each type of service for each shop. I categorize repairs into high-rated and low-rated repairs given the historical ratings that each repair has lead to for a given auto-repair shop. I then analyze how the distribution of these repairs changes according to a repair shop's rating state. I find that the closer a repair shop's rating is to surpassing a rounding threshold, the fewer low-rated repairs that the shop performs. This is evidence that shops are changing the services that they provide in response to ratings.

Second, I consider a firm's strategy of exerting extra effort in their repairs when their rating is close to a rounding threshold. A shop takes extra effort in order to make their customer happy and subsequently leave a five-star review. The technician can be extra friendly to the customer or better explain what is wrong with their car. The customer-service manager can offer the customer coffee. However, putting in this extra effort does have some cost for the shop, whether in time, money, or even mental exertion and therefore this behavior might not occur in every rating state. Since I cannot directly observe this strategy, I use a model to recover this behavior. I create a structural model to determine when it is optimal for a shop to engage in this behavior in order to maximize their future expected revenue. The structural model of firm strategy allows me to better inform the firms' managerial responses and can also point toward suggestions for optimal platform design. From the model, I can construct optimal policy functions describing when a shop should exert extra effort in order to maximize their ratings and thus expected future revenue.

Through these models I demonstrate the importance of how ratings can vary dynamically, and how firms' strategies change in response. Early in the review process, when a firm does not have many reviews, any single review can make a large impact in the average and thus displayed rating. Firms should be extra sensitive to their ratings when they have low numbers of reviews as any individual review has a larger impact on the average rating. The importance of this time frame is particularly true if early ratings have an impact on later ratings, as is further discussed in the literature review. Additionally, due to the rounding cutoffs on most review platforms, such as Yelp and the platform in my data context, when an average rating is close to these cutoffs, individual ratings also have larger impacts and the firm's incentives to obtain high reviews are even stronger.

In summary, this paper makes the following contributions. While existing research has gone to some length to examine how consumers respond to reviews, this is the first research to document previously unobserved short-run firm behavior. I utilize nonlinear incentives created by displayed-rating rounding thresholds to create a model of optimal firm behavior in order to improve their future ratings and thus future revenue stream. I show that firms do engage in a different distribution of behaviors, consistent with a greater incidence of selective, short-term actions designed to boost ratings and increase future revenue. I also use the model to consider counterfactuals in which ratings are rounded more (more discretization) and also not rounded. I find that firms exert more effort when the ratings are displayed at a more granular level. This suggests that consumer welfare can change depending on the way that ratings are displayed.

The paper proceeds as follows. Section 2 outlines some of the related literature. Section 3 explains a simple illustrative model to demonstrate the intuition and motivation of firms. Section 4 describes the data and Section 5 provides analysis for demand response to reviews. Section 6 provides descriptives on firms' response to reviews and Section 7 explores the two supply-side strategies. Section 8 outlines the theoretical model and Section 10 explores optimal policy functions, analyzes a counterfactual setting, and describes potential model extensions. Section 11 concludes.

2 Literature Review

This work contributes to several research streams in the marketing and economics literatures. Most notably this paper contributes to the literature on user-generated content and consumer reviews. Chevalier and Mayzlin 2006 is one of the first papers to document the causal impact of consumer reviews on demand. The authors compared book reviews across two platforms and found that positive reviews increased sales, and that reviews are overwhelmingly positive. As previously described, Luca 2016 used the displayed star rating rounding thresholds to look at the causal impact of ratings on revenue. Cabral and Hortacsu 2010, Zhu and Zhang 2010, and Janetos and Tilly 2017 all explore different ways that reviews affect consumers.

Rather than ratings per-say, another branch of the literature looks at other aspects of consumer feedback such as complaints or even tweets such as Knox and Oest 2014 and Ma, Sun, and Kekre 2015. Additionally, this work is closely related to the literature that examines certifications and official designations and how consumers and firms respond including Jin and Leslie 2003, Dranove and Jin 2010 and Fradkin et al. 2020.

While there is a large literature on how online reviews affect demand, supply-side reaction to reviews has been much less studied. This is an important question to consider and this paper seeks to help build the literature in this area. If firms are changing their response to reviews, then prior work looking at the demand side impact may actually be finding the combined effect of demand and supply. We need to understand both responses in order to isolate the effects of one.

There is an emerging literature on how firms respond to consumer reviews directly. For example Chevalier, Dover, and Mayzlin 2018 explore the effect of hotel managers responding to reviews by directly posting responses on TripAdvisor. However, rather than looking at the firm outcomes and actions, Chevalier, Dover, and Mayzlin 2018 explore how managers responding to reviews effects further review generation and content. Proserpio and Zervas 2017, Wang and Chaudhry 2018, and Gans and Lederman 2017 also look at how firms respond to reviews and the impact on subsequent reviews.

The most related papers to this work are Hollenbeck, Moorthy, and Proserpio 2019, Chen 2018 and Wang, Chaudhry, and Pazgal 2019. Hollenbeck, Moorthy, and Proserpio 2019 explore how hotels change their advertising strategy in response to their TripAdvisor ratings and find that ratings and advertising are complements. Chen 2018 looks at the impact of Yelp reviews on physicians. He finds an increase in demand and also finds that physicians call for more diagnostic tests once they are rated on Yelp, perhaps to please the patients and improve ratings. Wang, Chaudhry, and Pazgal 2019 explores whether quality actually improves due to ratings and finds that it does but that there is a heterogenous effect with branded chains versus independent firms. Yu, Debo, and Kapuscinski 2016, Kuksov and Xie 2010 and Chen and Xie 2005 explore pricing responses to reviews. The latter two papers are theoretical. To my knowledge, this is the first structural paper to analyze a firm's optimal strategy of short run actions, such as changes in services and effort, in order to improve ratings and thus future revenue stream.

Other papers in the consumer repair literature have explored how best to display and structure ratings on a platform such as Dai, Jin, and Lee 2018 and conducted textual analysis of the content of reviews to see what consumers deem most important, such as in Netzer et al. 2012.

This work is also related to the broad literature of reputation building and quality signaling for firms such as Fradkin et al. 2020. Additionally, the ratings discontinuities leads to nonlinear incentives for the firms. The nonlinear incentives relate to firm strategy broadly and can be seen in other settings such as branding premiums Klein and Leffler 1981 and salesforce incentives Misra and Nair 2011.

3 Stylized Model

First I will provide a story and intuition of what I believe is happening in this market which will inform my model. A firm's goal is to take on certain strategies in order to maximize the expected present discounted value of revenue based on dynamic rating. An auto-repair shop will have an underlying average rating and a displayed star rating on a platform. Based on the firm's current displayed rating, there is a probability that a customer walks into the auto-repair shop or not. Given their current rating state, (the average underlying rating and the number of ratings) a firm will decide whether to engage in certain costly actions to improve their next review. The shop receives revenue from that repair, π, minus the cost of their action, κ.

The key in the firm's decision to engage in costly effort is how likely any review is to move their displayed rating to a different half star by crossing a rounding threshold. Actions that the firm might take include turning consumers away who come in with repairs that have historically led to poor ratings or exerting extra effort to encourage a consumer to leave a 5-star review. The consumer will then decide whether or not to leave a review with probability p_r. If the consumer leaves a review, the shop's rating state changes, which affects their future revenue through the flow of consumers. Figure 2 demonstrates this process. The blue blocks are the firm and will be directly modeled. The grey blocks are the consumer and the consumers decisions are inputted into the model but not modeled themselves.

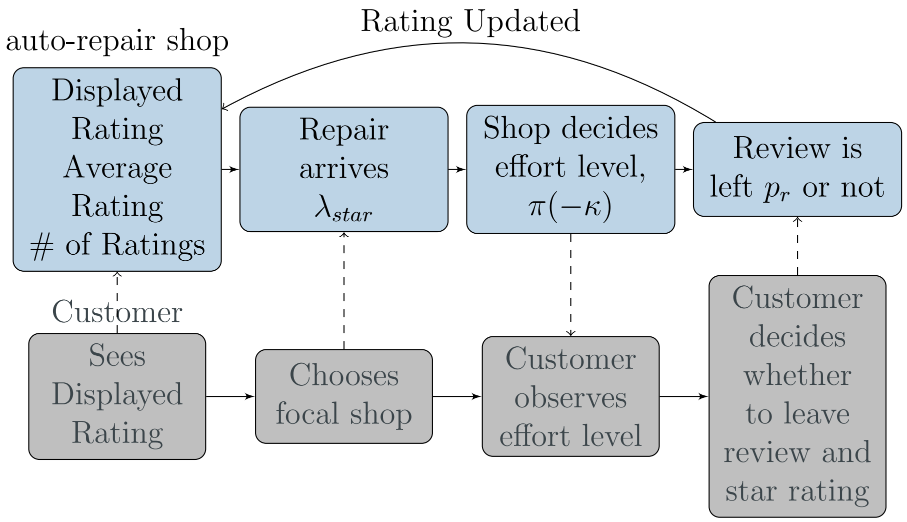

Rating Updated
auto-repair shop
Displayed Rating Average Rating # of Ratings
Repair arrives λ_{star}
Shop decides effort level, $\pi(-\kappa)$
Review is left p_r or not
Customer
Sees Displayed Rating
Chooses focal shop
Customer observes effort level
Customer decides whether to leave review and star rating

Figure 2: Model Flow Chart

3.1 Simple Example and Closed Form Solution

To get an idea of how important it is for firms to pay attention to and work to improve their ratings, I will demonstrate with a simplified example for which I can obtain a closed form solution. In this example, I consider that a firm takes on the strategy of turning away consumers if they believe the interaction will not lead to an optimal rating.

In this setting consider that a firm can take on one of three rating states.

- low rating (L)
- medium rating (M)
- high rating (H)

There are also two types of repairs, repairs that sometimes lead to low ratings, R_L and repairs that always lead to high ratings R_H, both of which give a shop the same revenue, π.

When an auto-repair shop is in the low rating state, they receive no repairs (no customers) and therefore the pay-off of being in that state is 0, V(L)=0. When a shop is in the high-rating state, they always receive a repair and it is always a high-rated repair. The value of being in this state in perpetuity is thus $V(H) = \frac{\pi}{1-\beta}$ where β is the shop's discount factor.

In the medium-rating state a shop also always receives a repair, however with probability p_L the repair is a low-rated repair and with probability $p_H = 1 - p_L$ it is a high-rated repair. If the shop accepts and performs a low-rated repair they obtain revenue π and with probability q they move down to state L and probability $1 - q$ they stay in their current state of M. If a shop obtains a high-rated repair they then move into state H in the next period.

A shops value function in state M is thus,

$$V(M) = p_H(\pi + \beta V(H)) + p_L(\max\{\beta V(M) + 0, \pi + \beta(q \cdot 0 + (1-q)V(M))\})$$

A repair shop in state M can have two strategies. Strategy A is to always accept low-rated repairs. Strategy R is to always reject low-rated repairs. If a shop uses strategy A then the shop's utility is:

$$V(A) = \frac{\pi[1 + p_H \frac{\beta}{1-\beta}]}{1 - p_L\beta(1-q)}.$$

For strategy R,

$$V(R) = \frac{p_H\pi[1 + \frac{\beta}{1-\beta}]}{1 - p_L\beta}.$$

It is unlikely that there are conditions where a shop should always accept or always reject, instead the strategy should vary depending on the current state. The percentage revenue that a firm loses in always accepting repairs in this state is $\frac{V(A)-V(R)}{V(R)}$. For illustrative purposes, I will consider a knife edge case to illustrate potential incentives and effects. Consider the case when $q = 0.5$ and $p_L = 0.5$, which means that the shop has a 50/50 chance of getting a high versus low-rated repair in state M and also a 50/50 chance of moving down a rating state if they accept the low-rated repair. If a shop accepts the low-rated repair in this case, they have a reduction in revenue by 33%! This is a large decrease and demonstrates how small changes in action can have large effects in a firm's revenue down the line.

4 Data and Context

4.1 Context

My context is the auto-repair industry. The auto-repair industry is an ideal setting to study my research questions for a variety of reasons. First, it is a highly information asymmetric market. Consumers do not know much about auto repairs. They often do not know what is wrong with their car, how much a repair should cost, or even if the repair was performed well ex-post. Therefore, it is likely to be a service that they will seek further information about before making a decision. Additionally, it is an expensive service. Consumers spend $850 annually on auto repairs or about 1% of their income[9], so again, consumers will most likely seek information before purchasing.

Furthermore, the auto-repair-service industry is an interesting and important market in of itself to study. It is quite a large market, with a revenue of $63 billion in the U.S. annually[10] and 229,000 auto-repair shops as of 2019[11]. This is also a highly fragmented industry: 75% of auto-repair shops are independent, as opposed to being chains or dealerships. We generally think of branding as being informative and helps to signal a certain quality and/or price level to consumers. Independent shops' quality are less known. The auto-repair shops in my data are independent mom and pop shops. They are not part of large chains or branded by the car make, and therefore building reputation through reviews is more important to them as they do not have a brand name to signal quality.

My data is transaction level data merged with consumer reviews. I have three sources of data. The transaction level data is from an established start-up in the auto service industry, which will henceforth be referred to as "the platform." The platform provides an online search tool for auto-repair shops that are deemed to be of a certain quality. An auto-repair shop can request to join the platform and will then go through an evaluation process

[9] Bureau of Labor Statistics, 2016

[10] https://aamcofranchises.com/research/how-big-is-the-industry/

[11] https://www.statista.com/statistics/436416/number-of-auto-repair-and--maintenance-shops-in-us/

to determine if the shop meets the platform's standards. As part of being approved and listed on the platform, an auto-repair shop has to provide the platform with their invoice data. They provide both their invoices going forward and invoices for at least a year prior to joining the platform. For each auto repair, I have information on the cost of the repair, the date of repair, an open text description written by the technician on the repair that was performed, as well as the make, model, year, and mileage of the vehicle. The data also contains the Vehicle Identification Number (VIN) which is a unique code and serial number that is used to identify each individual car. I also have shop level data which consists of an ID of the shop, the zipcode of the shop, and the date that the shop joined the platform (or was approved). On the consumer side I have the zipcode of the consumer, and whether or not they came to the shop through the platform or one of the third party partners (such as a towing service).

When a consumer visits the platform's website, they can search for a shop by location and find shops that are endorsed by the platform. In the resulting search listings, consumer ratings are shown and upon clicking on a particular auto-repair shop, individual text reviews and ratings are displayed. The ratings are displayed out of five stars rounded to the nearest half star.[12] The reviews are solicited from the platform from verified customers to these shops and are my second source of data. These survey responses can be matched to a specific invoice that was performed by the shop. For example, if a consumer came to a shop and got a brake pad replacement, I can then see the review they left after that particular repair. Since these reviews are solicited from verified consumers through the platform and not the shop, it is unlikely that they are manipulated or faked.

My final source of data is reviews from Yelp. As of December 2018, Yelp has 178 million unique visitors every month.[13] While Yelp is known most notably for restaurant

[12] The reviews are part of a longer survey that the platform sends to verified customers. One of the questions on the survey is would you recommend this shop on a scale of 0 to 10. This number is divided by 2 and used as the rating on the website. Therefore, unlike Yelp where a consumer has to leave 1,2,3,4 or 5 stars, the individual ratings can actually be a half star. There are other questions in the survey which I examine in Appendix Section C.

[13] `https://www.reviewtrackers.com/yelp-factsheet/`

reviews, 6% of reviews on Yelp are for the auto industry.[14] For every auto-repair shop for which I have transactional data, I have queried the shop in Yelp and scraped the resulting reviews including date, star rating, and text. Therefore I am able to reconstruct the displayed Yelp star rating for any time. For the top 350 zipcodes in my data, I have also scraped the first 5 pages of Yelp results for the query "auto repair" in order to obtain information about competition. I match a shop from Yelp to my data by first matching on zipcode and then doing a combination of fuzzy and manual matching on street address and shop name.

4.2 Data Descriptives

For the auto-repair shops in my data set I was able to find 1,388 on Yelp across 767 cities. I dropped invoices that were negative, zero, below $25 or above $10,000. Some of the negative invoices (and zeros) might be refunds, but others appear to be typos.[15] There are also a few quite huge numbers that seem to be typos as well. When looking at the number of invoices and number of new consumers I run a robustness test where I include these invoices as an invoice count (but not a revenue count).

On average, each shop provides about 3 years of invoice data, which gives me 10,324,552 invoices. The average invoice is about $400 as can be seen in Table 1. I also have 161,880 ratings that were solicited through the platform. Summary statistics of these data are provided in Table 2.

Table 1: Invoice Summary Statistics

Variable	**25%**	**Median**	**Mean**	**75%**	**Std Dev**
Invoice Total (2014$)	56.53	155.17	399.88	494.11	637.8
Number of Months of Data By Shop	30	42.3	41	50.8	18.5
Number of Invoices By Shop	3,271	6,196	8,087	10,551	7,517.9

From the Yelp data, I have 36,129 reviews for the shops for which I have transactional

[14] https://www.yelp.com/factsheet

[15] Since I cannot tell what is a refund I drop these from the analysis. There are also many zero invoice totals which I believe were just failure to include the actual invoice amount. Some of these drops should

data. As has been seen in other contexts, reviews are overwhelmingly positive. For an individual review the median rating is a 5 and the average is 4.37. At the end of my sample the average rating is 4.07 (displayed as 4 stars) with 26 reviews.

Table 2: Review Summary Statistics

Variable	**25%**	**Median**	**Mean**	**75%**	**Std Dev**
Individual Platform Rating	4.5	5	4.49	5	1.11
Individual Yelp Rating	5	5	4.37	5	1.35
Average Platform Rating	4.38	4.61	4.53	4.79	0.42
Average Yelp Rating	3.58	4.4	4.07	5	1.06
Number of Reviews Per Shop	5	12	26.04	29	51.96

Again, as has been documented before, reviews tend to be extremes on either end. There are more 1 star ratings than 2, 3, and 4 stars combined.

Table 3: Tabulation of Ratings

Star Rating	**Count**	**Percentage**
1	7,369	13.84
2	1,578	2.96
3	997	1.87
4	3,501	6.58
5	39,794	74.75

The top 10 repairs in my data set are: air conditioning, oil change, replace a part, flat tire, brakes, air filter, engine oil, light, lube, and diagnosis. After collapsing at the month level I see 3,863 displayed star changes, 2,028 of them are star changes up and 1,835 are star ratings down. This shows that shops' ratings are moving around quite often, which justifies that shops take actions as their ratings change.

For the model, I also need the rate at which reviews are left. To calculate this, I looked at how many new consumers a shop receives over their lifetime compared to how many reviews they receive on Yelp. The average rate of reviews being left is 1.3% with a standard deviation of 11.2%

5 Demand Side Response to Reviews

In order for firms to have an incentive to strategically respond to consumers in order to improve ratings, consumers must pay attention to and care about reviews. I first document that this is the case in my context. I aggregate my data at the monthly level. In the following regressions, an observation is a month, auto-repair shop, and rating. I sum all of the revenue that a repair shop has each month. I also take the cumulative average rating at the end of the month and what the rounded displayed rating is for that shop. Revenues are all inflation adjusted to 2019 dollars. I then de-mean the monthly revenue by shop to account for shops of various sizes. The second outcome variable I look at is the number of invoices that a shop has that month, again de-meaned at the shop level.

As it is unlikely that all consumers look up Yelp reviews and then immediately proceed to the auto-repair shop, I lag the ratings data one month.[16] For example, I take the rating at the end of October, and then look at the revenue obtained in November. It is likely that the time difference between the day of rating observation and the date of an invoice vary widely across consumers. When looking at various specifications of using contemporaneous reviews, the results remained qualitatively the same. The rating variable used in the following regressions is the displayed rating rather than the true average rating as that is what the consumer observes and thus is the information that is driving demand. A consumer can see every review on Yelp and therefore could construct the actual average rating themselves, however it is unlikely that a consumer will spend the time to do so, and even more unlikely that more than a small percentage of consumers do this.[17] The following regressions include city fixed effects. When a consumer is making a choice about which auto-repair shop to patronize, the consumer is likely making a choice within a certain geographic area. They are not comparing Bob's Auto Shop this month to Bob's auto shop last month. Therefore, city

reduce revenues and others increase it, so I assume on average it evens out.

[16] Table 11 in Appendix Section B runs the analysis without lagging the ratings data and the results are very similar.

[17] Most papers in the consumer review literature make this assumption as well.

fixed effects are chosen rather than shop fixed effects as those are the relative choices that consumers are making. Standard errors are clustered at the shop level.

Table 4 shows the output from one sample of these regressions. Output on other samples and specifications can be found in Tables 12 and 13 in the Appendix Section B and the results are similar. The results show that as displayed ratings increases one star, the monthly revenue for the shop increases by over $4000 and the number of invoices increases by over 8, which represent a 4.7% and 4.1% increase in revenue and invoices respectively.

Table 4: Ratings on Outcomes - Revenue and Number of Consumers

	monthly_revenue_demeaned	num_invoices_demeaned
	(1)	(2)
display_rating_lag1	4,118.78***	8.53***
	(1,151.22)	(2.72)
City Fixed Effects	Yes	Yes
Within Bandwidth	Yes	Yes
City Sample	2+	2+
Observations	10,302	10,302
Note:		*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level

The sample in these regressions include only auto-repair shops for which there is another auto-repair shop that is rated and listed on Yelp that month within the same city.[18]

Additionally, the sample is restricted to within a bandwidth of 0.1 around ratings. Because of this restriction, the shops on either side of the rating can now be considered more comparable. A common issue in ratings analysis is that there is an omitted variable bias of quality. Shops that are of a high quality are also likely to have high ratings and thus it is hard to disentangle what is driving the effect. However, to a certain extent, the rating is somewhat random depending on the type of consumer that a shop gets and who writes a review. For example, two shops of equal quality might have the same rating of 4.26, which

[18] A city seems like a reasonable competition area for which consumers might consider multiple shops. Zipcodes are small and arbitrary and MSA is too large.

gives a displayed rating of 4.5 stars. Then one gets a cranky customer who writes a bad review and their rating drops to a 4.24 and have a display of 4.0 stars. The other shop might get a particularly generous consumer who loves leaving 5-star reviews. Narrowing within the bandwidth gets closer to the causal estimates. However these results are still not causal and in the next section I run further analysis to get a causal interpretation.

5.1 New Consumers

New consumers are the ones who are most likely to be affected by ratings. If a consumer has already been going to an auto shop for awhile, it is unlikely they will continue to look up the reviews and change shops due to the reviews, rather they will decide based on their prior experience with the shop. New consumers, on the other hand, are more likely to look up and use reviews before deciding to go to a repair shop.

For close to 80% of my data I have valid VIN numbers which allow me to track cars over time. While most of the invoices have some input for VINs, a large percentage of them are invalid VINs.[19] This leaves me with 3,604,111 million unique VINs and over 8.2 million repairs. I make the assumption that tracking cars is the same as tracking a consumer or at least the drivers or family associated with that car. While it is possible I could be seeing the same car sold to a new owner, I will assume that a VIN represents a consumer, as it is unlikely that many of the cars are sold and then return to a shop that is in my data set. I repeat my analysis only for new consumers. To do this, I look at the Yelp displayed rating when the consumer first goes to a repair shop.[20] I then see how many new consumers the shop gets that month. I also construct a quasi-lifetime value measure for the consumer to see based on the rating they saw when they first chose that shop, how much revenue they sent to the shop up through the end of my data, as well as a quasi-lifetime number of visits

[19] I used the "check-digit validation" that checks that a VIN is valid for any vehicle sold in the U.S. or Canada. `https://en.wikibooks.org/wiki/Vehicle_Identification_Numbers_(VIN_codes)/Check_digit|`

[20] I might miss some new consumers if I do not have invoices that start before that consumer first went to the shop, but for most shops my data goes back several years before the reviews start so this should not

by summing across all their future visits. I will use these samples for the following regression discontinuity analysis.

5.2 Regression Discontinuity

The previous analysis demonstrates that reviews are associated with increased demand. In order to make a causal statement, I perform a regression discontinuity (RD) analysis. There are nine possible rounded ratings: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5 and 5.[21]. There are eight rounding thresholds: 1.25, 1.75, 2.25, 2.75, 3.25, 3.75, 4.25, and 4.75. For the regression discontinuity I consider the closest discontinuity to a particular rating. I then construct an indicator variable, *above cutoff* if the rating is above the closest discontinuity and therefore the rating is rounded up. For example if a shop has a rating of 3.15, the closest discontinuity is 3.25 and *above cutoff*, would be 0 as the displayed rating is 3 which is rounded down. A shop that has a rating of 4.8 has a closest discontinuity of 4.75. The *above cutoff* variable is 1 as the displayed rating is 5 which is rounded up.

For the RD, I normalized the rating with the closest rounding threshold so that I can compare across all rounding thresholds. Above is an indicator for whether or not the rating is rounded up. I used cross-validation to find the optimal bandwidth, although the results are mostly robust for bandwidths between 0.05 to 0.15, below 0.05 there is not enough data for power and larger than 0.15 it is less clear how comparable the shops are as the distance increases. As was also done previously, the outcome variables are demeaned at the shop level and fixed effects are still at the city level. Since only new consumers should care about the ratings, I perform this analysis on the new consumer data. The samples are the same as used in the OLS analysis, except in this case the top 1% of data is removed as there are some extreme outliers.

I consider three outcome variables, "CLV revenue" which is the amount of revenue

affect too many observations.

[21] A user has to leave a rating of 1,2,3,4, or 5 stars on Yelp

that a consumer brings in to the shop over the course of my data, the number of new consumers, and "CLV visits" which is the number of times a consumer returns to the shop over the course of my data.

The interpretation of the RD coefficients are as follows: for column (1) for a given shop, conditional on its rating being close to a cutoff, being above the cutoff is associated with a 16,173 dollar increase in revenue, after controlling for the true average rating. Conditional on the displayed rating, a higher true average rating is associated with lower revenue. However, for the normalized ratings within the bandwidth the standard deviation is 0.042 and the coefficient on normalized rating times one standard deviation is $-126,725 * 0.042 = -5299$. The addition of the two coefficients from this is positive at 10,874 which represents 12.5% increase in monthly revenue. Similarly for the other outcomes.

In order for the RD to be valid, there are a variety of validity tests to check. These can be found in Appendix Section D. There are also additional specifications and cross-validations on bandwidths.

Table 5: RD Various Outcomes

	CLV Revenue	Number New Consumers	CLV Visits
	(1)	(2)	(3)
normalized_rating_lag1	$-126{,}725.50^{**}$	-53.49	-340.03^{**}
	(64,128.15)	(32.70)	(142.93)
above_lag1	$16{,}173.94^{***}$	5.42^{**}	34.08^{***}
	(5,047.67)	(2.46)	(11.64)
City Fixed Effects	Yes	Yes	Yes
Bandwidth	0.08	0.08	0.08
Observations	7,727	7,727	7,727
Note:			*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level

The rest of my paper and my contribution to the literature explores how firms strategically respond to consumers in order to improve their future ratings. This might lead to some concern that the regression discontinuity analysis just presented is invalid. However, this should not be a concern. First, Lee and Lemieux 2010 state that:

> If individuals—even while having some influence—are unable to precisely manipulate the assignment variable, a consequence of this is that the variation in treatment near the threshold is randomized as though from a randomized experiment.

Shops cannot perfectly manipulate their rating, so the regression discontinuity is still valid, as there will still be shops on either side of the discontinuity that are trying to improve their ratings, as well as some who are not paying attention. Furthermore, the results could just be interpreted slightly differently, in that these are the causal demand effects of firm manipulated ratings.

6 Descriptive Evidence of Firm Behavior Change in Response to Reviews

Given that consumers care about and respond to reviews, firms should react to their reviews as well. In the next sections I present three pieces of evidence that firms are acting in order to change their ratings. Before I do so, I also wanted to provide anecdotal evidence that auto-repair shops are paying attention to their ratings.

During this research, I had conversations with several owners of auto-repair shops. I asked if they pay attention to their online ratings and they all said that they did. One repair shop owner stated: "First of all, I respond to all reviews good or bad.... I also have a service that tracks for reviews." Another said "You live and die by what folks say about you out there these days." The latter also mentioned that he paid attention to his ratings across different websites, in particular focusing on the platform from where my data comes from and Yelp; he looks at Google reviews to a lesser extent.

6.1 Review Incidence

Overall 1.3% of new consumers leave reviews on Yelp. However, the rate at which reviews are left varies depending on the shops' rating states. This indicates that the shops are doing something in order to increase the rate at which consumers leave reviews. In particular, the rate at which consumers leave reviews increases when the shop's rating is within a bandwidth[22] around the rounding threshold. Table 6 displays the changes in review incidence and what percent change over baseline that is, as well as the p-value of the level of significance of the difference of the first two columns. Within the bandwidth, 1.57% of consumers leave reviews, a 25% increase over baseline. Furthermore, more reviews are left when a shop's rating is just below the rounding threshold compared to just above. When looking at the 4.75 threshold, at which point a shop is rounded up to 5 stars or down to 4.5 if the rating drops below, the review rate differences are more extreme. For both this and the next piece of evidence, I looked at the shops' rating at the end of the week and then looked at the reviews that were left the corresponding week.

Table 6: Review Incidence in Various Rating States

Within Bandwidth	**Outside Bandwidth**	**Percent Increase**
1.57	1.25	25%***
Below Threshold	**Above Threshold**	**Percent Increase**
1.61	1.53	5%***
Within Bandwidth **4.75 Threshold**	**Outside Bandwidth** **4.75 Threshold**	**Percent Increase**
2.07	1.42	46%***
Below Threshold **4.75 Threshold**	**Above Threshold** **4.75 Threshold**	**Percent Increase**
1.88	1.54	22%***

*** p<0.01, ** p<0.05, * p<0.1

[22] bandwidth of 0.08, same as used in the RD

6.2 Average Rating

A second piece of evidence, is that the average rating that is left also changes depending on the firm's rating state. The average review left when within the bandwidth is higher than the overall reviews and higher than the average of reviews left outside the bandwidth.

Table 7: Average Review

Overall	**Within Bandwidth**	**Outside Bandwidth**
3.96	4.20	3.77

Difference between columns 1 and 2 and 1 and 3 significance*** p<0.01 level

6.3 Ratings Bunching

When plotting a histogram of ratings, there is a clear amount of bunching that appears just above rounding thresholds. In other words, there is an excess of mass of ratings right where a rating is rounded up to the next half star. There is also a dip in the distribution just to the left of the threshold. This can be seen visually in figure 3, which normalizes all ratings to the nearest rounding threshold.[23]

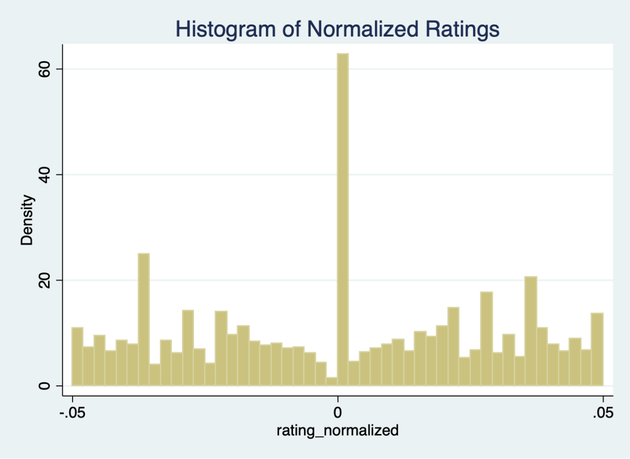


Figure 3: Histogram of Number of Ratings Normalized At All cutoffs

[23] For this analysis I used all of the shops for which I scraped Yelp reviews, so it not only includes the shops that I match to those for which I have invoice data, but also for their competitors in the same city.

We see that there is a large spike at 0, which represents all roundings thresholds. One might be concerned that with few reviews, due to the discreteness of the ratings, some ratings might occur more often than others. Figure 4 plots the ratings when there are at least 25 reviews, and the same pattern holds. Additionally, a dip just before the threshold can be seen as well.

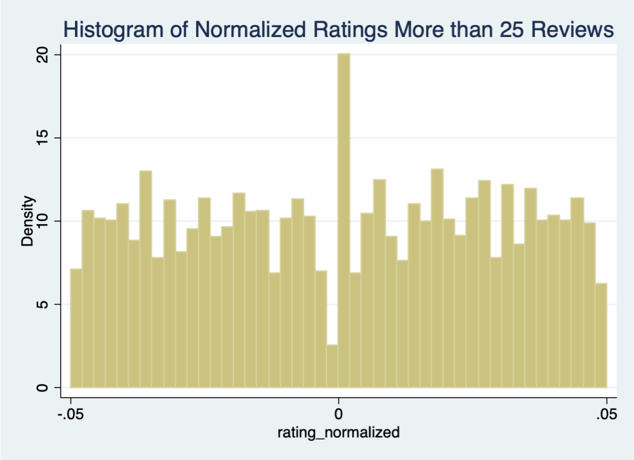


Figure 4: Histogram of Number of Ratings Normalized At All cutoffs For Shops with More than 25 Reviews

This pattern holds for each rounding threshold separately across all reviews as well as when I cut to shops that have even more reviews as can be seen in Appendix Section D.1.[24] There might also be a concern that this bunching comes from fake reviews. However, the pattern is the same in the survey data, which comes from verified customers and is very unlikely to be faked, as in figure 5. Additional Tables on the survey data follow similar patterns and can be found in Appendix Section D.1.

[24] Further evidence is shown in Appendix Section D.1 where I also looked at how these histograms change for shops that need only one, two, or three reviews to move across the threshold, as well as explored the pattern with different numbers of reviews.

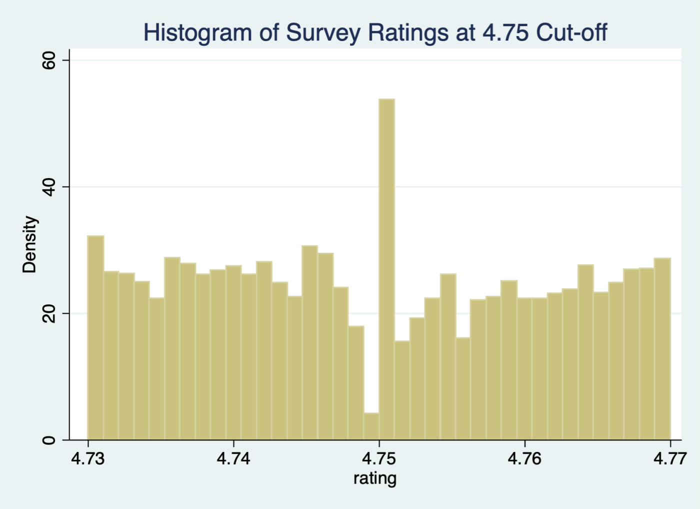


Figure 5: Histogram of Number of Ratings Normalized At All cutoffs Looking at the 4.75 Rounding cutoff Only

To ensure that this bunching is not only due to the discreteness of the data, I ran simulations of the ratings distributions. I took all the individual reviews that were given and for 2,000 shops. From the empirical distribution, I drew a number of reviews the shop had and drew subsequent ratings, with replacement. Essentially, I simulated what the distribution of ratings would look like if the order of the reviews was independent. The result is in Table 6. The blue is the real data and the pink is the simulated data. The real data includes 35% more mass just to the right of the threshold, and 32% less mass than the simulated data just to the left of the threshold. Therefore, while some bunching and troughs would occur naturally, there is an extra amount than would be expected.

To ensure that not only is this excess bunch and trough visibly apparent but also statistically significant, I use a Kolmogorov-Smirnov test to compare the simulated distri-

[25] In Appendix Section D.1, to test that this distribution is unlikely compared to a uniform distribution, I used Saez 2010's bunching method. In Saez 2010's method, one assumes that behavior at the threshold is a reasonable counterfactual for what behavior at the threshold would have been absent manipulation. The method is explained in the Appendix; I calculate the excess mass at the threshold looking at various size bins from 0.1 to 0.005 around the threshold and find $3.6 - 32.2\%$ excess mass, almost all of which is statistically significant using bootstrapped standard errors. For a bin size of 0.05, which is toward the larger ends of the 6 bins I used, I find an lack of excess mass. For a bin size of 0.0125, there are 9,892 observations and 2,895 shops around the threshold, which is 6.6% more mass than one would expect or 613 more reviews and 27 more shops which is economically significant as well. Therefore, I can reject the null that the distribution is smooth.

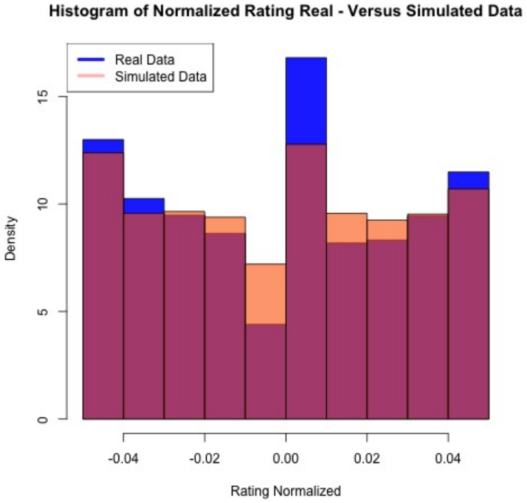


Figure 6: Histogram of Ratings Normalized to All cutoffs - Real Density and Simulated Data

bution in the absence of manipulation with the true empirical distribution.[25] I compare the distribution of the true and simulated data overall and obtain a p-value of 9.6e-14 indicating that the distributions are significantly different.[26] Additionally, I spilt the distribution below and above the threshold and compare them again, each time obtaining a p-value of $< 2.2e - 16$.[27] Another aspect to consider is which type of shops are in these bunches. This analysis can also be found in Appendix Section D.1.

All of this is evidence that firms are taking some strategic action in order to increase their number of ratings and their average ratings just enough to pass these displayed rounding thresholds. Once a firm is above the threshold, the consumer does not know (unless they bother to reconstruct averages) the difference between shop's averages, only that they are all the same star level. Thus, for example, having an average rating of 4.75, right above the threshold, looks the same to the consumer as having an average rating of 4.99, which is why the bunching is just right above the threshold. As a shop moves farther beyond the threshold there they have less incentive to strategically respond and therefore the shop pulls back in its strategic activity. In the following sections I will explore possible strategies that

[26] The D statistic is 0.044

[27] The D statistics are 0.075 and 0.12 for below and above the threshold respectively.

the firms take on in order to improve their ratings at the threshold.

7 Firm Ratings Manipulation Strategies

7.1 Rejecting Repairs

During my discussions with auto-repair shop owners, I also discussed whether or not they ever turn consumers away.[28][29] One owner said yes they do and it is because “Sometimes it’s that we can’t deliver in reality, what the client’s expectation is. Could be price-centric, time-centric, etc. . . ” Another said:

> When we [turn away jobs], it is not usually due to workflow [or capacity constraints] but more often in how our interaction is with the customer. We can tell from the start if it is going to be a bad match with us then there is no point in taking on the job from the beginning.

Consider the following example to illustrate how a shop might implement selectivity on type of service. Consider an auto-repair shop that believes their true quality is a 4.5-star shop. They currently have an average rating of 4.2, which is displayed as 4 stars. This is also just below the rounding threshold of 4.25 to be displayed as 4.5 stars. This shop only has seven reviews currently; thus any single 5-star review can push them over the threshold or a less than 5 star will keep bringing their rating down. A customer then walks into the shop with a repair that is not the shop’s strong suit and they have received a poor rating on that repair in the past. The shop might then refer the customer away. Whether or not an

[28] Evidence for this type of behavior has been seen in other markets as well. For example, in medicine it has been found that heart surgeons will refuse difficult operations in order to avoid poor mortality ratings. `https://www.telegraph.co.uk/science/2016/06/03/one-in-three-heart-surgeons-refuse-difficult-operations-to-avoid/` Although Yoon 2019 did not find this effect.

[29] I see some anecdotal evidence of this from consumer side data as well. I have a very small and noisy set of invoices (which I do not use in my analysis due to the size and lack of clean data) where I see consumers book appointments on the platform and leave messages for the repair shop. In several of them the consumer stated that the issue had been diagnosed by other shops before coming to the shop where they relieved the repair. While I don’t know if the consumer chose to walk away or the shop sent them away, this is further evidence that consumers sometimes leave the first shop they visit.

auto-repair shop refers the customer away should depend on their current rating state which is their average rating and the number of reviews they have.

I mention that the shop believes that they are a 4.5-star shop and wants their displayed rating to reflect what they believe their quality is. Whether or not a shop knows their true quality and the difference between their believed quality and their displayed quality is an issue that I address in the conclusion and leave to later research.

To see if this strategy is happening in practice, I first classify repairs and car make/model into low-rated and not low-rated for each shop. For car make and model the same story as above could happen, or the make or model of a car could indicate something about the consumer themselves. Additionally, the technician might have further unobservables that they use to make this decision that I as the econometrician cannot see. Since the median rating is 5 stars I count a repair as low rated if it did not receive 5 stars.[30] I know what repair was performed from a variable in the data which is the mechanic's notes on what they did in that invoice. This text variable is very messy. It is full of typos, shorthand, and various lingo, however I have cleaned this as much as possible. I then have created indicators for the top 213 repairs as classified by a mixture of repairs listed on the platform's website and the rate of words appearing in the text descriptions. Similarly for car make and model, I have cleaned as best as I can and in particular focused on the top 100 selling car make/models in the U.S.[31] I then run the following regressions:

$$\begin{aligned} \mathbf{1}_{\text{rating } < \text{ 5}} &= FE_{\text{top 213 repairs}} + FE_{\text{shop}} + FE_{\text{month date}} + \epsilon \\ \mathbf{1}_{\text{rating } < \text{ 5}} &= FE_{\text{top 100 car models}} + FE_{\text{shop}} + FE_{\text{month date}} + \epsilon \end{aligned} \tag{1}$$

The outcome, $\mathbf{1}_{\text{rating } < \text{ 5}}$ is an indicator for whether that repair was given a rating

[30] Unfortunately I do not observe ratings by repairs but rather by invoices. Most invoices contain 1-4 repairs. For example a repair might consist of a timing chain replacement, tire rotation, and an oil change. I only see the rating for this entire invoice. Therefore, I assign the rating to each repair in that invoice.

[31] As defined by focus2move,
`https://focus2move.com/category/best-selling-cars-ranking/usa-top-100/`

lower than a 5 star. These regressions give me a predicted probability of any repair not being rated a 5 for a given shop. For each month I sum the predicted probabilities across the month and average by the number of repairs performed[32]. This is an average percentage of low-rated repairs that a shop takes on that month.

$$Y_{jt} = \frac{1}{\text{num of repairs}_{jt}} \sum_{ijt} \hat{P}_{r<5,ijt}$$

$$= \text{percentage of invoices that month that contain a low-rated repair}$$

where $\hat{P}_{r<5,ijt}$ are the fitted values from (1). I use this as the outcome in the next regression.

$$Y_{jt} = \alpha + \beta_1 \bar{S}_{jt-1} + \beta_2 |\bar{S}_{jt-1} - C_{jt-1}| + \theta_j$$

$$\bar{S}_{jt} = \text{ Shop's average rating for that month}$$

$$C = \text{ cutoff points for where the rounded display rating changes, e.g. 4.25, 4.75}$$

$$\theta_j = \text{ Shop Fixed Effects}$$

As hypothesized, when a shop is above a rounding threshold, they take on fewer risky repairs/make and model combinations, however the further away from a threshold, the more risk they take on. The results are in Table 8. The interpretation is for every one standard deviation above the rounding threshold they move, they increase the amount of risk they take on by 17%.

[32] As mentioned, repairs and invoices are not the same. An invoice might have multiple repairs. Each repair is scored as high or low. The number of repairs performed that month, not number of invoices, is calculated by how many repairs are flagged, or 1 for "other" if none of the top 213 repairs are flagged.

Table 8: “Risky Repair” Regression

VARIABLES	Y_repair
yelp_rating	-0.0697***
	(0.000211)
disc_dist_abs	0.1726*
	(0.000187)
Observations	3,529
R-squared	0.7855
FE	Shop

SE Cluster at Shop
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Additionally, conducting the analysis in a different way, after a shop receives a rating of less than 5 stars, they take on fewer of those repairs in the future as seen in Table 9.

Table 9: Change in Number of Repairs

	(1)
VARIABLES	num_repairs_postrating
not5	-426.6***
	(43.07)
Constant	1,193***
	(38.11)
Observations	30,340
R-squared	0.151
Fixed Effects	City

Standard Errors Clustered at the Shop Level
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

7.2 Exerting Extra Effort

While there is anecdotal and statistical evidence that shops turn away certain repairs, and my toy model shows that in some settings this is profit maximizing, it is unlikely that all shops take part in this behavior or do so frequently. Another strategy for which firm’s might take to improve their ratings at the threshold is to exert extra effort towards their repairs and customers when they are close to improving their displayed rating. For now, I'm going to only consider temporary effort that is for one repair or customer only. I leave permanent quality investments and shifts to future research.

Some examples of the extra effort strategy might include, offering customers coffee, being extra friendly, or explaining the repair performed in detail. Shops might also take on actions to encourage the probability that a consumer leaves a review in general if they feel that the consumer had a good experience; for example, providing consumers with coupons if they send in screenshots of a five-star review[33] or informing the consumer of how important reviews are for their small local business. Since this extra effort is unobserved, I cannot provide similar descriptive evidence of this behavior. Therefore, I create a structural model to back-out this strategy, which I describe in the following section.

8 Model

Consider a single firm with no competition and a firm's inherent quality is fixed over time. The state space consists of a shop's true average rating and the number of reviews that a shop has.

State Space: $x_{it} = \{S_{it}, N_{it}\}$

- S: the firm's current true rating average
  - $S \in [0, 5]$, continuous
  - $\bar{S}$ is the rounded displayed rating and takes on these values: $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$
- N: number of reviews
  - $N \in [0, 1, 2, 3, ...125]$[34]

[33] I personally have experienced this in a variety of markets from purchasing goods on Amazon to tours during travel. I have also heard stories of this happening to others such as in home improvement services.

[34] 97.5% of shops have fewer than 125 reviews, so I cut the max number of reviews here. Additionally, any one review does not change the average rating much at this point.

I aggregate the data up to the weekly level. A time period is thus a week.[35] I chose a week, because it seems reasonable that a repair shop owner would check their ratings each week and make decisions at the weekly level. Each week the shop receives a certain number of repairs based on their displayed rating from the end of the previous week, as that is what the consumer sees.[36] The higher a shop's rating, the more repairs they receive. The number of repairs that a shop receives each week is $R(S)$ which is a Poisson distribution with mean λ where λ is calibrated from empirical distribution in the data for each rating state as follows:

$$\lambda(S) = \begin{cases} 32, & \bar{S} = 1 \\ 34, & \bar{S} = 1.5 \\ 36, & \bar{S} = 2 \\ 38, & \bar{S} = 2.5 \\ 40, & \bar{S} = 3 \\ 42, & \bar{S} = 3.5 \\ 44, & \bar{S} = 4 \\ 46, & \bar{S} = 4.5 \\ 48, & \bar{S} = 5 \end{cases}$$

Each period, the firm has to decide whether or not to exert extra effort, with cost κ, on all repairs they receive that week, or not. The action space is thus:

$$a = \{ \text{ exert normal effort, exert extra effort with cost } \kappa \}$$

These actions are unobserved to the econometrician.

[35] Since the model is at the weekly level, I re-do my main regression specifications at the weekly level which can be seen in Tables 14 and 28 in Appendix Section D.

[36] Unfortunately, as was true in the descriptive analysis, I cannot know when a consumer viewed the rating. Some consumers may view a rating online and head right to the auto-repair shop. Others may look up the reviews and go to the shop days or weeks later. A week level aggregation seems reasonable to allow a lag of when the consumer goes to the shop.

A firm's flow utility function is their revenue. The current period revenue is:

$$\pi(a) = \begin{cases} \pi & \text{if exert regular effort} \\ \pi - \kappa * R(S) & \text{if exert high effort} \\ 0 & \text{if no repair arrives} \end{cases}$$

Upon receiving a repair, the customer might choose to leave a review. The probability of a review being left is calibrated from the data and is $p_r = 1.75\%$. The rating that is left will depend on whether or not the shop exerted extra effort. Since two, three and four stars are left very infrequently, it is hard to identify changes in these star ratings with and without effort. Therefore, for ease of estimation, I fix the probabilities of these ratings as is calculated from the data. Let the probabilities of receiving a particular star rating be $\{p_{1star}, p_{2star}, p_{3star}, p_{4star}, p_{5star}\}$. If a shop exerts extra effort then the probabilities of receiving a 1 and a 5 star are $\{p_{e,1star}, p_{e,5star}\}$.[37] There is no change with the other star ratings. I estimate the probability that a 1 star rating is left with no effort as well as the change in probability with effort, Δ. I can then recover the probability of 5 stars with and without effort since the probabilities sum to 1.[38] Thus,

$$p(\bar{e}) = \{p_{1stareffort}, p_{2star}, p_{3star}, p_{4star}, p_{5stareffort})$$

$$p(\underline{e}) = \{p_{1star}, p_{2star}, p_{3star}, p_{4star}, p_{5star})$$

Ratings are drawn from a multinomial distribution, one with effort and one without. The draws are s, or s_e if extra effort is provided. The number of new ratings that are received in a given week are,

$$M(R(S)) \sim Binomial(R(S), p_r)$$

The individual ratings are drawn and averaged from the following distribution:

[37] Where $p_{2star} = p_{e,2star}$, $p_{3star} = p_{e,3star}$ and $p_{4star} = p_{e,4star}$

[38] So $p_{1star} = \Delta + p_{e,1star}$ and $p_{e,5star} = \Delta + p_{5star}$.

$$A(M(R(S)),e) \sim Multinomial(Mdraws, p(e))$$

where $p(e)$ is the probability of getting each star rating depending on the effort level. The state transition equations are thus:

$$S' = \frac{N}{N+M(R(S))}S + \frac{M(R(S))}{N+M(R(S))}A(M(R(S)),e)$$

$$N' = N + M(R(S))$$

Let $\bar{e}$ be extra effort and $\underline{\text{e}}$ be normal effort. The firm exerts effort if and only if

$$((\pi - \kappa) * R(S) + \beta E[V(S'(\bar{e}), N')]) - (\pi * R(S) + *\beta E[V(S'(\underline{\text{e}}), N')]) > 0 \tag{2}$$

9 Identification and Estimation

9.1 Identification

Let e be the amount of effort. Let $\underline{e}$ be regular effort and $\bar{e}$ be extra effort. I need to estimate $p_1(e)$ in order to see how effort translates into ratings. This is how the probability of getting a one star rating varies by the amount of effort put in. Additionally, I need to estimate e, how a shop decides effort depending on their rating state, in order to know how effort varies by state. The firm will choose to exert effort deterministically according to equation 2.

I could also consider that the firm chooses to exert effort with

$$(\pi - \lambda * \kappa + \beta V(S'(\bar{e}), N')) - (\pi + *\beta V(S'(\underline{\text{e}}), N')) + \epsilon > 0$$

where the support of ϵ is bounded. A possible extension of the model is to consider that ϵ has infinite support, however I will not be pursuing this version.

More specifically, I am estimating three parameters, the probability of getting 1 star with no effort p_{1star}, the change in probability of obtaining 1 star with effort, Δ, and the cost of effort κ. This allows me to obtain $p_{1star,effort} = \Delta + p_{1star}$ and the probability of getting a five star with and without extra effort is thus just 1 minus the sum of the rest.

The parameters are identified as follows. There are going to be some rating states in which a firm would never exert high effort as there is no point, the benefit is not worth the cost. The average rating in these states will be a baseline which allows me to identify p_{1star}. There are other states in which a firm would almost certainly exert effort for reasonable costs of effort. An assumption I am making here is that the error term has bounded support. If the error term had infinite support, then there would never be states in which a firm would always exert extra effort as long as there is some non-zero cost of effort. For example, there could be some shock or error that brings in a particularly difficult customer, in which even if the shop is in a state in which extra effort should be exerted, they will not be able to or want to.

By looking at the average ratings received in these states compared to the states in which no effort should ever be exerted, I can identify Δ. Finally, by looking at the average rating overall, this allows me to see the cost of effort. The higher the cost of effort is, the fewer states in which effort will be performed. As long as Δ is positive, the overall average rating is a decreasing function of κ. Once Δ and p_{1star} are pinned down, I can then move around κ until I hit the overall average rating as they are monotonically related. I use the same moments described in the data in Section 6, the average rating left within a bandwidth around a threshold, outside the bandwidth, and overall.

Writing this more formally:

$$P(1star|state) = P(1star|effort)*P(effort|state)+P(1star|noeffort)*(1-P(effort|state))$$

Assumption 1: There exists a state S, N such that the probability of one star given

S, N is $p_{1star,noeffort}$.

Assumption 2: There exists a state S', N' such that the probability of one star given S', N' is $p_{1star,effort}$.

Assumption 3: Let $F(\kappa, p_{1star,noeffort}, \Delta)$ be the function that maps parameters to the overall average rating from a model simulation from the model solution given those parameters. For any $p_{1star,noeffort}, \Delta$, the function $F(\cdot, p_{1star,noeffort}, \Delta)$ is strictly monotonically decreasing in κ and surjective.

Under these assumptions the model is identified. The fraction of one star reviews in state S, N identifies $p_{1star,noeffort}$. The fraction of one star reviews in state S', N' identifies $p_{1star,noeffort} + \Delta$ and thus identifies Δ. Given those parameters, and Assumption 3, for any average rating there is exactly one κ such that $F(\kappa, p_{1star,noeffort}, \Delta)$ equals the average rating.

Suppose that instead of effort versus no effort, a shop chooses a continuous level of effort, $e \in [0, 1]$ for a cost $e * \kappa$, and a benefit $e * \Delta$. The same argument applies. Assumption 3 is loosely equivalent to to saying that the average level of effort $e \in [0, 1]$ decreases in κ. This also generalizes to the cost being any function, $F(\kappa, e)$ as long as Assumption 3 holds.

Let us consider another thought experiment for the identification. First, assume that there are no ratings. Therefore, there is no reason for a shop to exert extra effort (assuming that this does not affect repeat purchases). Then $p_{1star,noeffort}$ is identified. Now consider that there are ratings, but they are not publicly displayed to consumers. There is still no need for the shops to "game" the ratings. Now consider the current state of the world, with displayed ratings and rounding thresholds. Consider if a shop's average rating is as far from a rounding threshold as it can be (for example an average rating of 4, which is 0.25 from the 3.75 rounding threshold and the 4.25 rounding threshold). A shop that has 100 reviews in this state has little incentive to exert effort because they are unlikely to get close to a threshold. Looking at the average rating here helps to identify $p_{1star,noeffort}$. Then consider a shop with a rating of 4.24. This shop has a lot of incentive to exert extra effort because they are very close to having their displayed rating change. Every shop should exert extra effort in this state so this identifies Δ by comparing the average rating in this state compared to the average rating in the previous described state. Next consider the state of a shop at 4.15 average rating, which is further away from the threshold but not that far. Shops with lots of ratings might not have an incentive to exert effort at this rating state, and shops with few ratings have an incentive to exert effort, as they have the chance to move past the threshold. By identifying the point at which a shop will move from exerting no effort to effort in terms of the number of reviews, this can identify the cost of effort, κ.[39]

9.2 Estimation

I estimate the model using a two-step process. First, I solve the model. From the model solution, I obtain an optimal policy function stating in which states the firm should exert extra effort. I then assume that the firms follow this optimal policy. That way I can construct whether or not the firm exerted effort in this state which I then call "observed" effort. This is similar to some sales force models, such as Misra and Nair 2011 in which the relevant action, which is effort put out by the sales force in their case, by the firm in mine, is unobserved to the econometrician and has to be inferred from sales for them and by optimal policy and average ratings for me. Misra and Nair 2011 translate sales policy function to an "effort policy function." I perform a similar process.

I then estimate the probabilities of 1 star with no effort and the difference in 1 and 5 stars with and without effort, Δ. Fixing these ratings probabilities, I loop through various values of κ and find the one that minimizes the distance between the simulated and the true moments. The revenue of each repair is normalized to 1 and thus the cost of effort can

[39] As a final test to the model being identified, after performing Monte Carlo simulations I was also able

be thought of as a percentage of the revenue. Given my estimated cost of effort, I repeat the process. I re-solve the model with the new parameter estimates, find the optimal policy function assuming these parameters, and next find the probabilities. I repeat this process until the model converges. More details of the estimation can be found in Appendix Section F.

10 Optimal Policy Functions and Counterfactuals

10.1 Optimal Policy Functions

One output of the model is to obtain optimal policy functions that describe in which states the firm should exert extra effort. I can then see how much firm value would change if the firm never took on extra effort. When a firm has 5 reviews and has an average rating of 4.75 (they are right at the cutoff to be displayed as 5 stars rather than 4.5), the firm would lose 24.5% of firm value by never turning away risky repairs. When a firm has more reviews, the change is not as extreme, but still important. At 50 reviews the firm would lose 1% of firm value.

In the extra effort model, at the same state, 4.75, with 5 reviews by sometimes exerting extra effort a firm can increase firm value by 3.9%, with 50 reviews 8.6%, with 120 reviews 11.5%.

Figure 7 is the percentage of states in which effort is exerted as a function of the number of reviews according to the optimal policy. The rate at which effort is exerted decreases the more reviews a shop has because any one review will make less of a difference to the average rating. The rate increases again at the end due to the finite nature of the model. A shop is forward looking and wants to end with a high rating.

to recover the parameters.

Figure 7: Percentage of Rating States in Which Effort Is Exerted

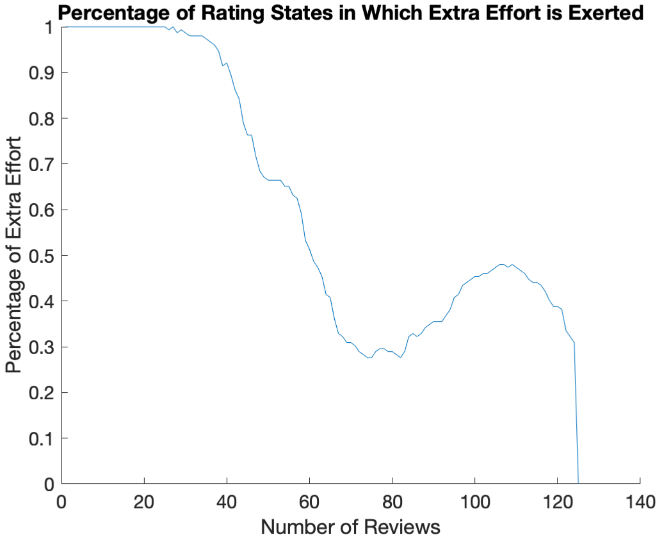
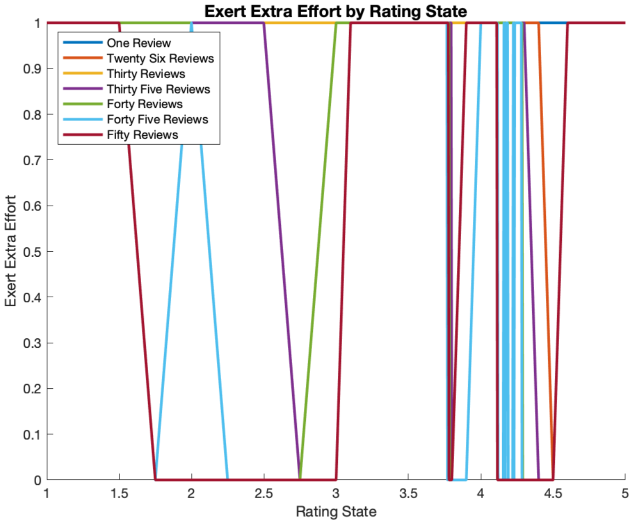


Figure 8 shows when extra effort is exerted depending on the average rating. Each line represents a different number of reviews that the shop might have. The more reviews a shop has, the less often effort is exerted again. Also, the area in which effort is not exerted tends to be when the rating is far from the rounding thresholds. The top represents effort being exerted and the bottom is no effort exerted. The exertion of effort bounces around at the rounding thresholds. The fewer reviews the more it bounces, and the more reviews the larger space in which no effort is exerted.

Figure 8: Optimal Policy - Exerting Effort By Rating State and Number of Reviews



10.2 Counterfactuals

Using the model I consider a variety of counterfactual ways in which ratings might be displayed.

10.2.1 More Rounding

For any counterfactual, an important note to consider is what assumptions am I making or holding fixed. I first consider the counterfactual world where roundings are more discrete, and ratings are displayed to the nearest whole star rather than half.[40] In this counterfactual I am assuming that the cognitive response with consumers is the same as the half star displays. Here the displayed ratings are 1, 2, 3, 4 or 5 stars. The rate of repair arrivals is thus,

[40] As has been common in the movie industry for decades.

$$\lambda(S) = \begin{cases} 28, & \bar{S} = 1 \\ 34, & \bar{S} = 2 \\ 40, & \bar{S} = 3 \\ 47, & \bar{S} = 4 \\ 53, & \bar{S} = 5 \end{cases}$$

I find that firms' incentive to exert extra effort decreases, and that effort decreases by 25% in the observed-rating states.

This points to rounding being detrimental to consumers. There are nuances to consider, such as if extra effort goes down, fewer customers receive this improved service. Additionally, the reviews might reflect the extra effort, and therefore a consumer does not get what they expect.

This begs the question of why do platforms round reviews to begin with? There are several possible behavioral explanations. One is that perhaps consumers mentally round anyway. There is extensive literature on left digit bias in which consumers only pay attention to the leftmost digit and round in their head anyway, whether with car mileage (Lacetera, Pope, and Sydnor 2012), or prices, (Anderson and Simester 2003). There is also evidence Donkor 2019 that providing a reduced menu reduces the cognitive load on consumer decision-making.

10.2.2 No Rounding

I also considered the counterfactual world where ratings are not rounded (or rather rounded to the nearest 0.1) and the average star rating is displayed instead. An assumption here is the consumers are not mentally rounding, and they are in fact paying attention to the average rating. In this world there are not certain thresholds or cutoffs where firms should care more or less about their ratings, rather they just always want to obtain the highest rating possible. To see what happens in this scenario, rather than having the number of repairs that a shop obtains be determined by the displayed rating, there is a continuum of how many repairs arrive based on the average rating. Thus, rather than 8, repair arrival looks like

$$\lambda(S) = \begin{cases} 9, & \bar{S} = 0 \\ 10, & \bar{S} = 0.1 \\ & \vdots \\ 50 & \bar{S} = 4.9 \\ 51, & \bar{S} = 5 \end{cases}$$

When ratings are not rounded, firms have an incentive to exert extra effort more often, and they exert effort 30% more of the time in the observed-rating states. In the world where there are rounding thresholds, firms only have an incentive to exert extra effort when their rating is close to a rounding threshold. If they are far from a rounding threshold, then there is not much that can be done to change their displayed review. However, when reviews are not rounded, every review changes the displayed rating (when the number of reviews total is not too large and depending on the number of digits displayed), therefore firms actually exert more effort. In both scenarios, the more reviews a shop has, the less helpful effort is. In this way, rounding ratings may be hurting consumers, in that they are not getting the best quality service they can. Furthermore, in this counterfactual world, the average rating increases from 3.96 to 4.21, which could increase firm revenue, although this depends on how consumers respond to the non-rounded ratings and also increases competition.

I also made the rounding even less discrete, to the 0.01, and found that effort increased 40.1% compared to the current displayed ratings.

10.2.3 Everyone Leaves a Review

Next I considered a counterfactual where everyone leaves a review. Several platforms and other papers have considered ways to encourage consumers to leave reviews, but is this actually helpful?

I changed the max number of reviews to be 250, and this is hit within 6 weeks. Effort is observed to happen 62.8% more of the time. Therefore, it seems that if more consumers leave reviews, shops are more likely to engage in the extra effort behavior.

10.2.4 More Recent Reviews Are Weighted More Heavily

Some theoretical papers, such as Vellodi 2020, have suggested that more recent reviews be weighted more heavily in order to lower barriers to entry of new firms. This is also a good idea in order for firms to continually work on their effort and quality and not "rest on their laurels." In order to run this counterfactual, I weight reviews from the current week as contributing $\frac{1}{\alpha}$ to the total average rating and reviews from the previous week are weighted $\frac{1}{\alpha^2}$ etc. I choose this method to make the estimation easier (as this way the state space does not increase too much). The number of reviews no longer matters.

I find the following results. When $\alpha = 0.5$, shops exert effort in 98.7% of states which is a 94.3% increase from the original result.

Table 10: Varying Alpha

α	**Percentage of observed states with effort**	**Percentage change in effort from baseline**
1	98.7%	94%
0.2	98.7%	94%
0.15	95.2%	87%
0.1	46%	-10%
0	0	-50%

11 Conclusion

Given that consumers pay attention to ratings which then means that ratings have a causal effect on revenue and number of consumers, firms should also pay attention to ratings. Due to the way that ratings are displayed on most platforms, this creates incentives for firms to act change their behavior when their rating is close to moving past a rounding threshold. Using novel data which allowed me to match reviews to invoices and using a structural model to back out unobserved behavior, this is the first paper to document such short-run strategic actions on behalf of the firm.

There are several considerations to pursue in future related research. Namely, this work begs the question, what is the optimal way to display ratings to align consumer and firm incentives and improve welfare? Why did this star rounding become the standard way to display ratings? It could be that consumers already mentally round. Star ratings could also reduce the mental load on consumers by making comparisons easier (similar to reducing options with 401(k) plans and menu options with tipping).

Other future questions include, how do long run actions change due to ratings platforms? For example, do firms change their quality levels? In this paper I assumed a fixed level of quality, but rather than short run extra effort actions, a firm could make long run quality investments. Additionally, do the ratings rounding create more homogeneity in how consumers see firms than exists in reality? Further exploring the downsides to displaying rounding ratings is also important. Shops may not exert extra effort or a few idiosyncratic shocks could destroy a high-quality shop's reputation.

References

Anderson, Eric T. and Duncan I. Simester (2003). "Effects of $9 Price Endings on Retail Sales: Evidence from Field Experiments". In: *Quantitative Marketing and Economics*

1.1, pp. 93–110. ISSN: 15707156. DOI: 10.1023/A:1023581927405. URL: http://link.springer.com/article/10.1023/A:1023581927405.

Cabral, L and Ali Hortacsu (2010). “The dynamics of seller reputation - theory and evidence from eBay”. In: *Journal of Industrial Economics* LVIII.1, pp. 1–58. URL: papers2://publication/uuid/361E7193-F45B-4AD8-A2DE-C65CA0BD5AA5.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik (2014). “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs”. In: *Econometrica* 82.6, pp. 2295–2326. ISSN: 1468-0262. DOI: 10.3982/ecta11757.

Chen, Yiwei (2018). “User-Generated Physician Ratings-Evidence from Yelp”. In: URL: https://www.softwareadvice.com/resources/how-patients-use-.

Chen, Yubo and Jinhong Xie (2005). “Third-Party Product Review and Firm Marketing Strategy”. In: *Marketing Science* 24.2, pp. 218–240. ISSN: 0732-2399. DOI: 10.1287/mksc.1040.0089.

Chevalier, Judith A., Yaniv Dover, and Dina Mayzlin (2018). “Channels of Impact: User Reviews When Quality Is Dynamic and Managers Respond”. In: *Marketing Science.* ISSN: 0732-2399. DOI: 10.2139/ssrn.2766873.

Chevalier, Judith A. and Dina Mayzlin (2006). “The Effect of Word of Mouth on Sales: Online Book Reviews”. In: *Journal of Marketing Research* 43.3, pp. 345–354. ISSN: 0022-2437. DOI: 10.1509/jmkr.43.3.345. arXiv: 0022-2437.

Dai, Weijia, Ginger Jin, and Jungmin Lee (2018). “Aggregation of Consumer Ratings : An Application to Yelp . com”. In: *Quantitative Marketing and Economics* 16.289-339, pp. 289–339. ISSN: 0021-9886. DOI: 10.1111/1468-5965.00396. URL: http://www.hbs.edu/faculty/PublicationFiles/13-042{_}3c56fcb5-b7b9-4eb9-a026-2822c79b7127.pdf.

Donkor, Kwabena B (2019). “How Difficult is Tipping ? Using Structual and Non-Structrual Approaches to Estimate Decision Costs”. In: *Working Paper.*

Dranove, David and Ginger Zhe Jin (2010). "Quality disclosure and certification: Theory and practice". In: *Journal of Economic Literature* 48.4, pp. 935–963. ISSN: 00220515. DOI: 10.1257/jel.48.4.935.

Farronato, Chiara et al. (2019). "Consumer Protection in an Online World: An Analysis of Occupational Licensing". In: *Working Paper*. URL: https://www.ftc.gov/about-ftc.

Fradkin, Andrey et al. (2020). "CONSUMER PROTECTION IN AN ONLINE WORLD : An Analysis of Occupational Licensing". In: *NBER Working Paper* 26601. URL: http://www.nber.org/papers/w26601.

Gans, Joshua S and Mara Lederman (2017). "Exit, tweets and loyalty". In: *NBER Working Paper*.

Hahn, Jinyong, Petra Todd, and Wilbert Van Der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design". In: *Econometrica* 35.1, pp. 189–190.

Hartmann, Wesley, Harikesh S. Nair, and Sridhar Narayanan (2011). "Identifying causal marketing mix effects using a regression discontinuity design". In: *Marketing Science* 30.6, pp. 1079–1097. ISSN: 1526548X. DOI: 10.1287/mksc.1110.0670.

Hollenbeck, Brett, Sridhar Moorthy, and Davide Proserpio (2019). "Advertising Strategy in the presence of reviews: An empirical Analysis". In: *Marketing Science*.

Janetos, Nick and Jan Tilly (2017). "Reputation Dynamics in a Market for Illicit Drugs". In: March, p. 1. arXiv: 1703.01937. URL: https://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/reputation-dynamics{_}0.pdf.

Jin, Ginger and Phillip Leslie (2003). "Quality : Evidence From Restaurant Hygiene". In: *Quarterly Journal of Economics* May, pp. 409–451.

Klein, Benjamin and Keith B. Leffler (1981). "The Role of Market Forces in Assuring Contractual Performance". In: *Journal of Political Economy* 89.4, pp. 615–641.

Knox, George and Rutger van Oest (2014). "Customer Complaints and Recovery Effectiveness: A Customer Base Approach". In: *Journal of Marketing* 78.September, pp. 42–57. ISSN: 0022-2429. DOI: 10.2139/ssrn.1427265.

Kuksov, Dmitri and Ying Xie (2010). "Pricing, Frills, and Customer Ratings". In: *Marketing Science* 29.5, pp. 925–943. ISSN: 0732-2399. DOI: 10.1287/mksc.1100.0571.

Lacetera, By Nicola, Devin G Pope, and Justin R Sydnor (2012). "Heuristic Thinking and Limited Attention in the Car Market". In: *American Economic Review* 102.5, pp. 2206–2236.

Lee, David S. (2008). "Randomized experiments from non-random selection in U.S. House elections". In: *Journal of Econometrics* 142.2, pp. 675–697. ISSN: 03044076. DOI: 10.1016/j.jeconom.2007.05.004.

Lee, David S. and Thomas Lemieux (2010). "Regression Discontinuity designs in economics". In: *Journal of Economic Literature* 48.2, pp. 281–355. ISSN: 00220515. DOI: 10.1257/jel.48.2.281.

Luca, Michael (2016). "The Case of Reviews , Reputation, and Revenue : The Case of Yelp.com". In: *HBS Working Paper*.

Ma, Liye, Baohong Sun, and Sunder Kekre (2015). "The Squeaky Wheel Gets the Grease—An Empirical Analysis of Customer Voice and Firm Intervention on Twitter". In: *Marketing Science* 34.5, pp. 627–645. ISSN: 0732-2399. DOI: 10.1287/mksc.2015.0912.

Misra, Sanjog and Harikesh S. Nair (2011). "A structural model of sales-force compensation dynamics: Estimation and field implementation". In: *Quantitative Marketing and Economics* 9.3, pp. 211–257. ISSN: 15707156. DOI: 10.1007/s11129-011-9096-1.

Moe, Wendy W. and David A. Schweidel (2012). "Online product opinions: Incidence, evaluation, and evolution". In: *Marketing Science* 31.3, pp. 372–386. ISSN: 07322399. DOI: 10.1287/mksc.1110.0662.

Moe, Wendy W. and Michael Trusov (2011). "Measuring the Value of Social Dynamics in Online Product Ratings Forums". In: *Journal of Marketing Research* XLVIII.June, pp. 444–456. DOI: 10.2139/ssrn.1479792.

Netzer, Oded et al. (2012). "Mine Your Own Business: Market-Structure Surveillance Through Text Mining". In: *Marketing Science* 31.3, pp. 521–543. ISSN: 0732-2399. DOI: 10.1287/mksc.1120.0713.

Proserpio, Davide and Georgios Zervas (2017). "Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews". In: *Marketing Science*. ISSN: 0732-2399. DOI: 10.2139/ssrn.2521190.

Saez, Emmanuel (2010). "Do Taxpayers Bunch at Kink Points?" In: *American Economic Journal: Economic Policy 2* 2.August, pp. 180–212.

Tadelis, Steven (2016). "Reputation and Feedback Systems in Online Platform Markets". In: *Annuel Review of Economics*, pp. 321–342. DOI: 10.1146/annurev-economics-080315-015325.

Talwar, Arjun, Radu Jurca, and Boi Faltings (2007). "Understanding user behavior in online feedback reporting". In: *Proceedings of the 8th ACM conference on Electronic commerce*, p. 134. DOI: 10.1145/1250910.1250931.

Vellodi, Nikhil (2020). "Ratings Design and Barriers to Entry". In: *Working Paper*, pp. 1–54. ISSN: 1556-5068. DOI: 10.2139/ssrn.3267061.

Wang, Yang and Alexander Chaudhry (2018). "When and How Managers' Responses to Online Reviews Affect Subsequent Reviews". In: *Journal of Marketing Research* LV.April, pp. 163–177. ISSN: 0022-2437. DOI: 10.2139/ssrn.2831402.

Wang, Yang, Alexander Chaudhry, and Amit I. Pazgal (2019). "Do Online Reviews Improve Product Quality? Evidence from Hotel Reviews on Travel Sites." In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3238510.

Yoon, Tae Jung (2019). “Quality Information Disclosure and Patient Reallocation in the Healthcare Industry: Evidence from Cardiac Surgery Report Cards”. In: *Marketing Science* January 2020. ISSN: 0732-2399. DOI: `10.1287/mksc.2018.1146`.

Yu, Man, Laurens Debo, and Roman Kapuscinski (2016). “Strategic Waiting for Consumer-Generated Quality Information: Dynamic Pricing of New Experience Goods”. In: *Management Science*. DOI: `10.2139/ssrn.2334423`.

Zhu, Feng and Xiaoquan (Michael) Zhang (2010). “Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics”. In: *Journal of Marketing* 74.2, pp. 133–148. ISSN: 0022-2429. DOI: `10.1509/jmkg.74.2.133`. arXiv: `/doi.org/10.1509/jmkg.74.2.i [https:]`.

A Appendix

B Demand Robustness Checks

First, I provide additional specifications and samples for the OLS demand side analysis. The first robustness check is to see how not using the lagged displayed rating but rather the contemporaneous rating affects the results. The reason would be if consumers went into a shop immediately after reading Yelp reviews. The results are very similar showing they are not sensitive to this timing assumption as seen in Table 11.

Table 11: Ratings on Outcomes - Unlagged Reviews

	monthly_revenue_demeaned	num_invoices_demeaned
	(1)	(2)
display_rating	3,962.47***	8.43***
	(1,159.07)	(2.74)
City Fixed Effects	Yes	Yes
Within Bandwidth	No	Yes
City Sample	2+	2+
Observations	10,302	10,302

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

Next, I run the OLS regression restricting the sample one step at a time. In both Tables 12 and 13 the first column is on the unrestricted sample, i.e. all shops and cities. The second column is the same specification as in the main text for comparison. In this sample, I limited to cities which have at least two shops in my sample (i.e. cities with competition for which ratings might be more relevant) and for ratings that are within the bandwidth around the ratings threshold to start to get closer to a causal estimate. Columns (3) and (4) add in controls for the number of reviews. The number of reviews does not have a statistically significant impact, but the coefficient is positive and larger than the standard error.

Table 12: Ratings on Outcomes - Revenue

	monthly_revenue_demeaned		
	(1)	(2)	(3)
display_rating_lag1	787.39*	899.29*	4,118.78***
	(424.20)	(459.09)	(1,151.22)
City Fixed Effects	Yes	Yes	Yes
Within Bandwidth	No	No	Yes
City Sample	All	2+	2+
Observations	40,792	30,923	10,302

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

Table 13: Ratings on Outcomes - Revenue

	num_invoices_demeaned		
	(1)	(2)	(3)
display_rating_lag1	0.79	0.89	8.53***
	(1.04)	(1.10)	(2.72)
City Fixed Effects	Yes	Yes	Yes
Within Bandwidth	No	No	Yes
City Sample	All	2+	2+
Observations	40,792	30,923	10,302
Note:			*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level

Finally, I run the regression at the weekly level, as that is the level of time aggregation I will use in the model.

Table 14: Ratings on Outcomes - Weekly

	weekly_revenue_demeaned	num_invoices_demeaned
	(1)	(2)
display_rating_lag1	905.31***	1.82***
	(252.96)	(0.61)
City Fixed Effects	Yes	Yes
Within Bandwidth	Yes	Yes
City Sample	2+	2+
Observations	44,131	44,131
Note:		*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level

C Survey Questions on Demand

In addition to the average rating of the shop and how that affects demand, the platform also asks a variety of other questions to consumers about their repair experience. I looked at how these correlate with demand outcomes. For this analysis I restricted my data to invoices that occurred from consumers going through the platform. The sample size is smaller as not many consumers come through the platform, but the number of shops is higher as I did not have to restrict to those I matched on Yelp.

If consumers rate the car as being ready when promised as opposed to not ready, the shop gets $2,755 more in revenue and 9.45 more repairs per week, although this rating is not significant.

Table 15: Was Your Car Ready? Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Was your car ready? Rating	2,755.92	9.45
	(2,330.22)	(6.39)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 16: Quality of Work - Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Quality of work Rating	1,483.77**	3.00
	(683.42)	(1.97)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 17: Did you pay a fair price? - Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Price Fair? Rating	864.37**	0.97
	(405.08)	(1.17)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 18: Did you feel pressured to get more repairs? Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Feel pressured? Rating	−529.15	−7.22
	(2,958.24)	(8.28)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 19: Was the work explained to you? Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Work explained? Rating	4,949.06**	18.48**
	(2,288.53)	(8.00)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 20: Was the work done correctly on first try - Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Correctly first time? Rating	3,084.08*	8.74*
	(1,598.56)	(4.76)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 21: Were you treated with Respect - Outcomes

	weekly_revenue	weekly_num_of_repairs
	(1)	(2)
Respect? Rating	2,197.31	9.49
	(2,159.98)	(6.36)
Month Fixed Effects	Yes	Yes
Shop Fixed Effects	Yes	Yes
Number of Shops	1709	1709
Observations	200,258	200,258
Note:		*p<0.1; **p<0.05; ***p<0.01

Summarizing these regressions on the survey questions it seems that the most important qualities that create benefits for shops are, quality of work , the work being explained to the consumer, the consumer feeling like they got a fair price, and the overall rating a consumer leaves.

D Regression Discontinuity: Validity and Robustness

I checked the RD at varies bandwidths. While it is not significant for all bandwidths, it is significant at most except the extreme bandwidths and the coefficients are still the proper direction.

Table 22: CLV Outcome - RD at Various Bandwidths

	invoice_clv_demeaned				
	(1)	(2)	(3)	(4)	(5)
normalized_rating_lag1	509,337.00	−267,237.10**	−141,616.30	−81,847.54	−91,749.28**
	(2,533,084.00)	(130,712.00)	(117,027.70)	(78,708.44)	(38,120.23)
above_lag1	3,592.55	22,796.11***	19,104.76**	12,325.72	15,954.52**
	(27,265.48)	(8,399.50)	(8,382.26)	(8,088.95)	(7,155.85)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes
Bandwidth	0.01	0.05	0.08	0.1	0.15
Observations	1,635	6,115	7,806	12,270	16,986

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

Table 23: CLV Visits Outcome - RD at Various Bandwidth

	num_visits_clv_demeaned				
	(1)	(2)	(3)	(4)	(5)
normalized_rating_lag1	3,451.66	−540.36*	−328.79	−286.13*	−230.24**
	(7,395.25)	(295.27)	(287.66)	(169.32)	(94.81)
above_lag1	−27.88	47.11**	41.20*	29.26	33.19*
	(77.76)	(19.16)	(21.68)	(19.00)	(18.21)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes
Bandwidth	0.01	0.05	0.08	0.1	0.15
Observations	1,635	6,115	7,806	12,270	16,986

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

Table 24: Number of New Customers - RD at Various Bandwidths

	newcon_num_demeaned				
	(1)	(2)	(3)	(4)	(5)
normalized_rating_lag1	1,169.15	−74.08	−41.95	−54.98**	−20.70
	(1,420.26)	(54.57)	(41.55)	(25.23)	(16.25)
above_lag1	−13.09	6.27*	5.65*	3.81	2.77
	(14.19)	(3.36)	(3.31)	(3.10)	(2.77)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes
Bandwidth	0.01	0.05	0.08	0.1	0.15
Observations	1,635	6,115	7,806	12,270	16,986

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

I also used cross validation using optimal MSE to find a bandwidth of 0.046 and the results are in table 27. [41]

Table 25: RD with CV Bandwidth

	CLV Revenue	Number New Consumers	CLV Visits
	(1)	(2)	(3)
normalized_rating_lag1	−149,275.10	−28.55	−285.92
	(200,066.50)	(117.00)	(501.37)
above_lag1	20,246.87**	5.26	42.92*
	(9,893.86)	(5.30)	(23.80)
City Fixed Effects	Yes	Yes	Yes
Bandwidth	0.046	0.046	0.046
Observations	4,502	4,502	4,502

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

In the traditional regression discontinuity design, a treatment is given to all participants at the same time. However, in my setting firms move across the "treatment" threshold at a variety of times. There might be a concern that the affect of Yelp is changing over time. Therefore, I re-ran the RD splitting the sample across time. I split the sample into the pre-2018 data and the 2018 data. The magnitudes of the coefficients are similar, although they are more statistically significant in the latter period, perhaps because Yelp became more mainstream, but more likely because the sample is larger. The coefficients are in the right direction and larger than the standard errors for the pre-2018 data, but the sample is much smaller.

[41] Calonico, Cattaneo, and Titiunik 2014

Table 26: RD Various Outcomes

	CLV Revenue		Number New Consumers	
	Pre-2018	2018	Pre-2018	2018
	(1)	(2)	(3)	(4)
normalized_rating_lag1	−87,831.52	−56,487.84	−13.20	-66.31^{*}
	(143,925.70)	(91,603.50)	(58.79)	(34.33)
above_lag1	9,178.45	$10{,}233.56^{**}$	5.76	4.37
	(9,057.41)	(5,150.89)	(4.35)	(2.92)
City Fixed Effects	Yes	Yes	Yes	Yes
Bandwidth	0.08	0.08	0.08	0.08
Observations	1,966	4,140	1,966	4,140
Note:			*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level	

Table 27: RD Various Outcomes

	CLV Visits	
	Pre-2018	2018
	(1)	(2)
normalized_rating_lag1	−163.63	-310.70^{**}
	(266.75)	(155.06)
above_lag1	25.46	27.35^{**}
	(15.63)	(11.80)
City Fixed Effects	Yes	Yes
Bandwidth	0.08	0.08
Observations	1,966	4,140
Note:		*p<0.1; **p<0.05; ***p<0.01 Standard Errors Clustered at the shop level

I also ran the RD at the weekly level as that is the level of time I will use in the model.

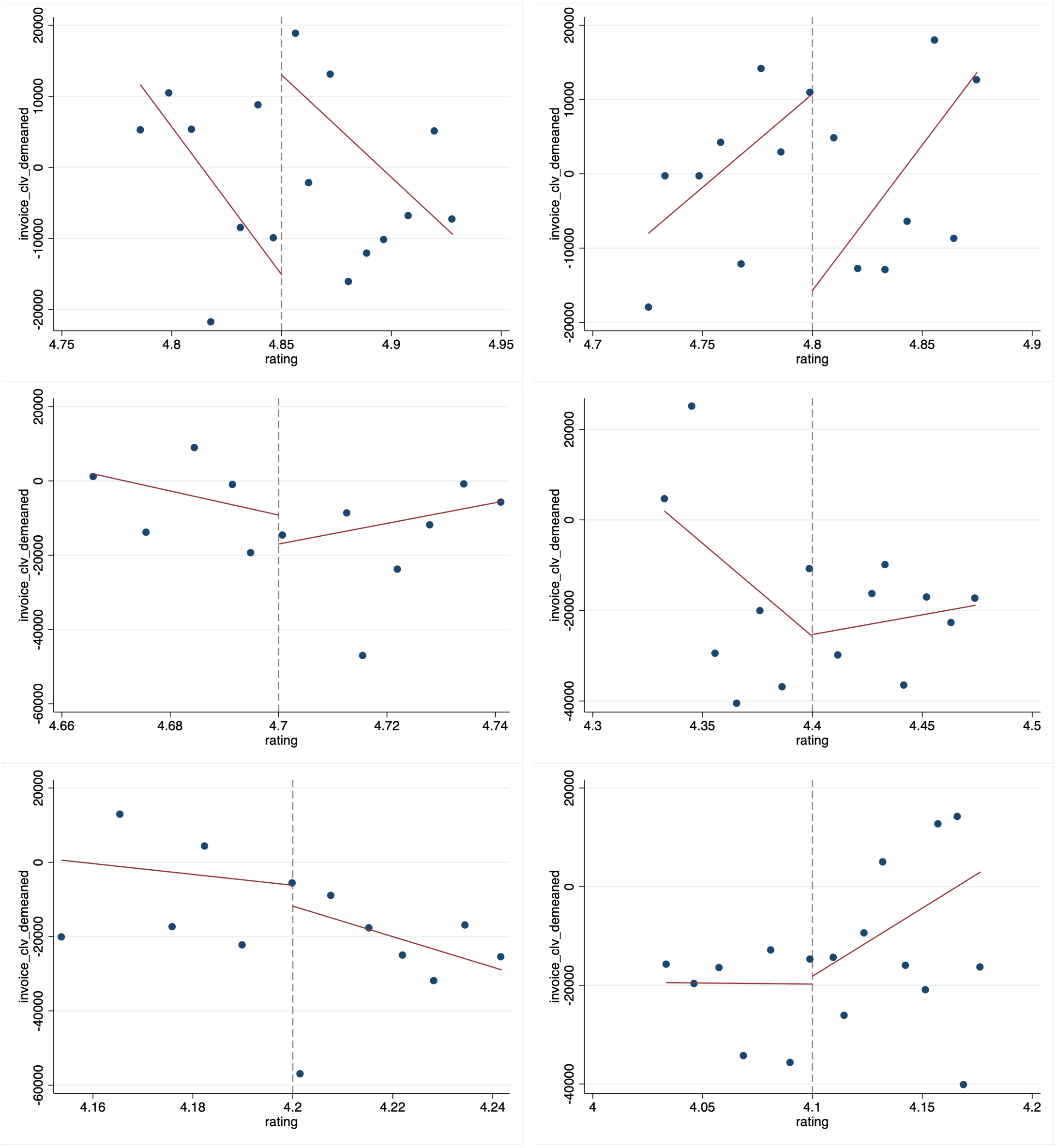
Table 28: RD Various Outcomes - Weekly

	CLV Revenue	Number New Consumers	CLV Visits
	(1)	(2)	(3)
normalized_rating_lag1	$-56{,}487.99^{***}$	-7.90	-126.99^{**}
	(20,370.59)	(9.72)	(52.95)
above_lag1	$5{,}452.83^{***}$	1.00	12.13^{***}
	(1,669.00)	(0.73)	(4.29)
City Fixed Effects	Yes	Yes	Yes
Bandwidth	0.08	0.08	0.08
Observations	32,980	32,980	32,980

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors Clustered at the shop level

As a placebo test, I also checked for discontinuities at random points rather than at the rounding threshold. There does not appear to be a clear pattern.

Figure 9: Placebo Test Discontinuities at Random Thresholds

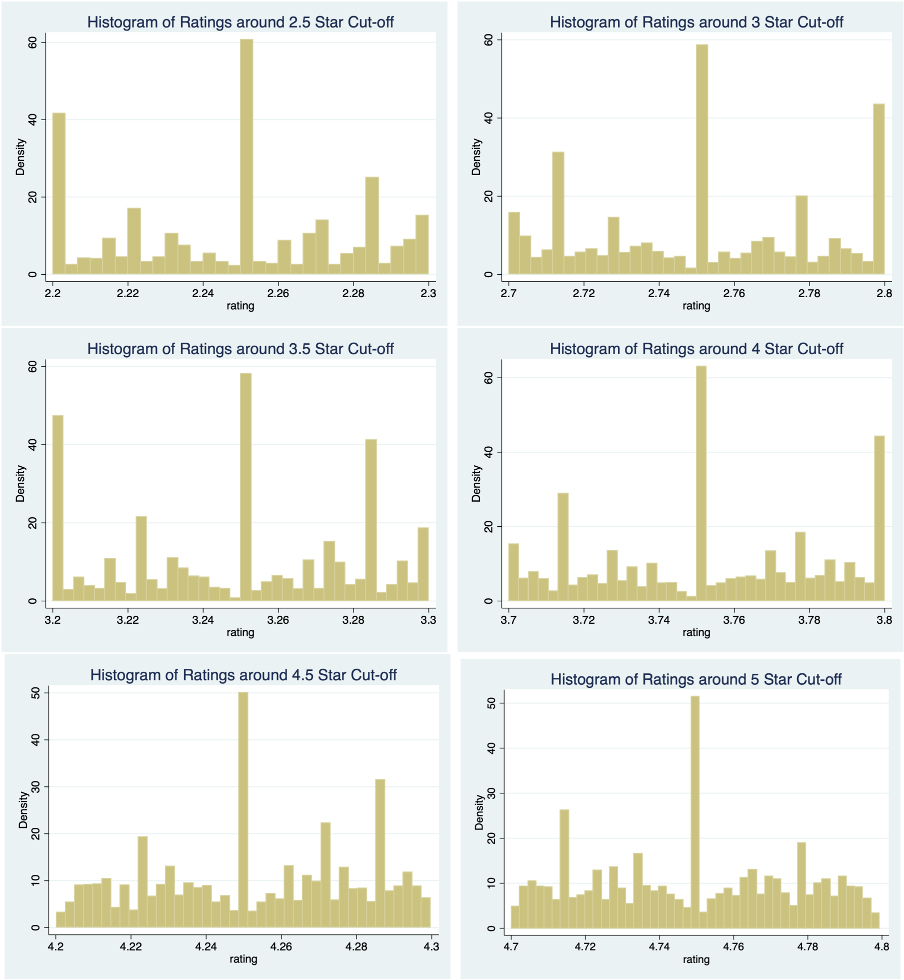


D.1 Bunching

D.1.1 Additional Bunching Graphs

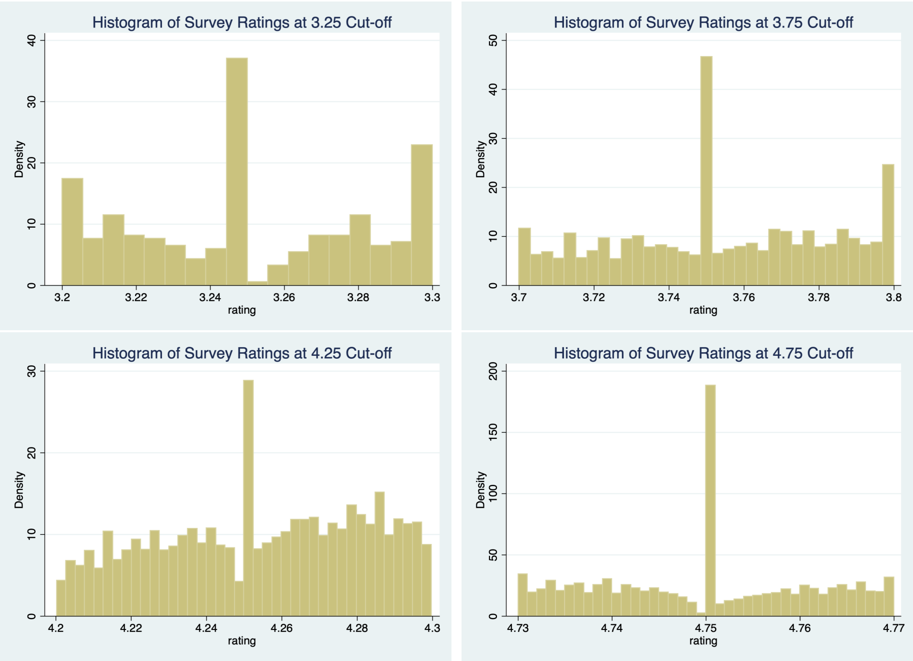
In addition to checking the bunching across all rounding thresholds at once, I looked at the distribution for each rounding threshold separately. Bunching can be seen for each.

Figure 10: Bunching at Each Rounding Threshold



Since Yelp data is subject to fake reviews, I also looked at bunching in the survey data conducted by the platform. Since the platform collects this data from verified consumers, it would be very difficult to have fake reviews in this data. Bunching can be seen across each rounding threshold in this rating data as well.

Figure 11: Bunching Across Thresholds Using Platform Survey Data


Histogram of Survey Ratings at 3.25 Cut-off

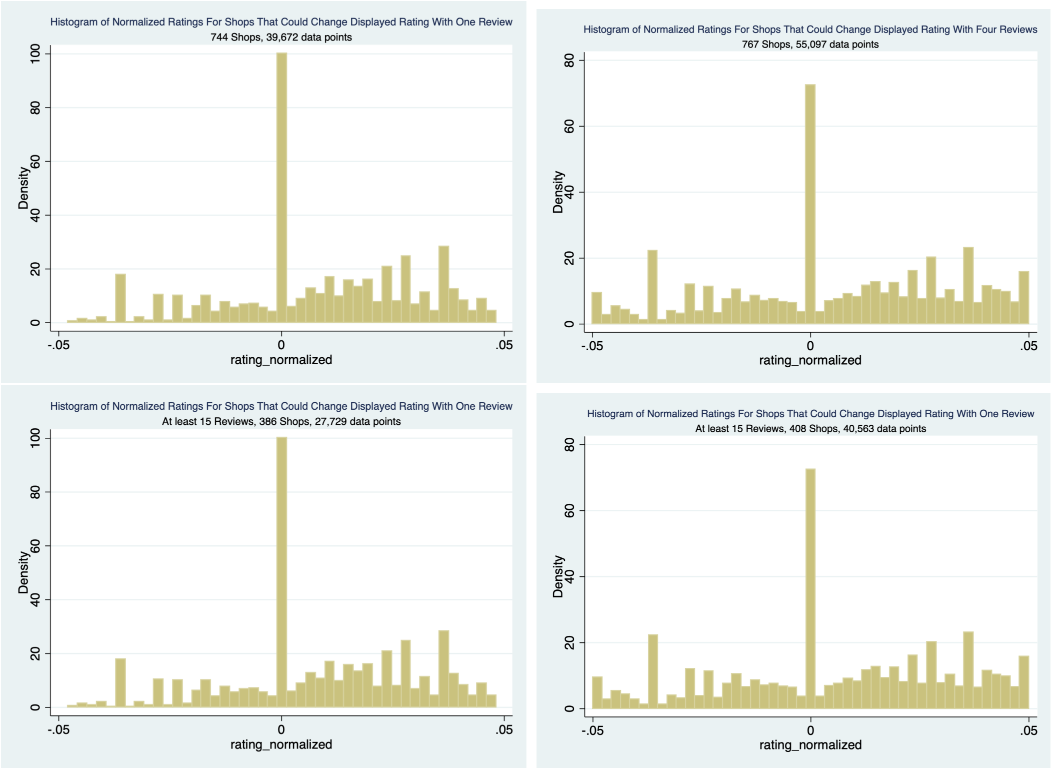
Histogram of Survey Ratings at 3.75 Cut-off

Histogram of Survey Ratings at 4.25 Cut-off

Histogram of Survey Ratings at 4.75 Cut-off

It is also interesting to consider how shops might behave when they only need a few reviews to move past the rounding threshold. I plotted the data only for shops whose displayed rating could change with 1 or 4 reviews, as is seen in the first row. In the second row I used the same data but cut the sample to shops that have at least 15 reviews already. We see the bunching is even more extreme in these samples.

Figure 12: Bunching For Shops That Only Need A Certain Number of Ratings To Move Past A Threshold



D.2 Bunching Analysis

As described in the main text, in order to check for the statistical significance of bunching, in addition to the Kolmogorov-Smirnov test I use bunching estimation by Saez 2010. For a variety of different binwidths (δ), at threshold z, I calculate the excess bunching at the threshold, B.

$$B = \int_{z^*-\delta}^{z^*+\delta} h(z)dz - \int_{z^*-\delta}^{z^*-2\delta} h(z)dz - \int_{z^*+\delta}^{z^*+2\delta} h(z)dz$$

Where

$$\hat{h}(z^*)_+ = \hat{H}^*_+/\delta \qquad \hat{h}(z^*)_- = \hat{H}^*_-/\delta$$

$$\hat{B} = \hat{H}^* - (\hat{H}^*_+ + \hat{H}^*_-)$$

My estimates are as follows, where the first column is for all ratings I have data for and the second column is only for ratings associated with the shops in my main data set.

Table 29: B at Different Bin Sizes

B	**B Platform Shops Only**	**Bin Width**
-0.100	-0.13	0.15
0.075	0.040	0.1
0.036	0.079	0.075
-0.048	0.0299	0.05
0.056	0.043	0.025
0.066	0.097	0.0125
0.322	0.315	0.005
0.610	0.561	0.0025
0.896	0.825	0.001

E Which firms bunch?

To explore the bunching phenomena more, I looked at the timing between ratings as is displayed in table 30. First, on average, there are 77 days between reviews. However, after getting a one star review, the time to the next review (both median and mean) increases significantly ($p<0.001$). On the other hand, when looking at the time between a one star review and the next review, if that next review is a 5 star review, the time decreases significantly ($p = 0.02$). Perhaps this is because the shops are making an extra effort to get that review to bump them back up.

Table 30: Days Between Reviews

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
overall	0	8	26	76.68	76	3086	1623
after1star	0	10	34	93.58	97	2280	329
after1starto5	0	8	31	87.2	88	2280	NA

Next I compared shops that are ever in a “bunching” spot as well as looked at the months in which they are in the “bunch.” For both comparisons, the shops that bunch tend to be larger, both in terms of revenue and number of consumers.

Table 31: Comparing Stores That Ever Bunch and Those That Do Not

bunch_ind	monthly_revenue	newcon_num	num_of_invoices_monthly	n
0	74900.6	66.73	180.1	46664
1	95358.81	91.47	230.95	50528

Table 32: Comparing Bunching Months with Non Bunch Months

bunch_ind	monthly_revenue	newcon_num	num_of_invoices_monthly	n
0	83401.48	76.94	201.25	96191
1	94901.33	90.97	227.29	1283

Finally, comparing overall summary statistics, the shops that bunch tend to have fewer reviews on average.

Table 33: Overall and At Bunch

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
overall_numreviews	1.000	9.000	24.000	55.108	60.000	750.000
bunch_numreviews	4.000	4.000	16.000	35.180	36.000	281.000
overall_rating	1.000	3.879	4.407	4.199	4.741	5.000
bunch_rating	1.250	3.750	4.250	4.091	4.750	4.760
overall_rating_var	0.000	0.556	1.510	1.564	2.500	4.000
bunch_rating_var	0.182	0.644	1.688	1.636	2.688	3.938
overall_rating_norm	-0.250	-0.125	0.002	0.009	0.150	0.250
bunch_rating_norm	0.000	0.000	0.000	0.002	0.002	0.010

F Model Estimation

In order to estimate the model on my data I take the following steps.

1. Simulate transition matrices.
    - First I simulate data to estimate the probability of moving to any state from any state, where a state is made up of average rating and number of reviews.

2. Solve the Model.

    - Next I solve the model using value function iteration. This also gives me optimal policy functions of in which states firms should exert extra effort.

3. Set initial guesses for parameter values.

4. Simulate data based on the initial guesses.

5. Back out "observed effort" from the simulated data. I assume that firms act according to the optimal policy functions.

6. Calculate the moments from the true data.

7. Calculate the moments in the simulated data.

8. Calculate the distance between the true moments and the simulated moments.

9. Update the parameter values and re-simulate the data, re-calculate moments, until the simulated moments and the true moments converge.

G Future Model Extensions

In future work, I hope to extend the model in many ways.

Increase Review Incidence

First, an alternative story, is that in addition to (or instead of) the review distribution shifting due to effort, effort could instead increase the probability that a review is left. I demonstrated evidence of this in section 6. Second, is to consider inherent quality of the auto-repair shop. I will sort firms into high and low quality shops and see how their strategic behavior differs. For example, should a low quality firm exert more in order to pretend to be high quality, or is it better to have one's true quality apparent? This model could also be extended to include a continuum of quality levels.

Next, I plan to create a model with two types of effort. One which is a temporary effort in the current time period and one which is a longer term investment. Enough of these longer term investments could move a shop from low to high quality.

Further minor extensions to the model include:

1. Allow the number of reviews to influence the probability of a shop receiving a repair.

2. Multiple (or two) types of repairs. For example, hard repairs and easy repairs, which would have different costs of effort and reviews would be drawn from different ratings distributions.

3. Allow the probability of a consumer writing a review to vary with the current rating state of the shop, the type of repair performed, and whether or not the shop exerted extra effort.[42]

[42] Previous literature has shown that consumers are more likely to write a review when their opinion differs from the previous few reviews as in Moe and Schweidel 2012.