The Consequences of Rating Inflation on Platforms: Evidence from a Quasi-Experiment

Arslan Aziz
Sauder School of Business, The University of British Columbia, arslan.aziz@sauder.ubc.ca,

Hui Li
HKU Business School, The University of Hong Kong, huil1@hku.hk,

Rahul Telang
H. John Heinz III College, Carnegie Mellon University, rtelang@andrew.cmu.edu,

Informative online ratings enable digital platforms to reduce the search cost for buyers to find good sellers. However, rating inflation, a phenomenon in which average rating increases and rating variance across listings decreases, threatens the informativeness of ratings. We empirically identify the consequences of rating inflation by conducting a quasi-experiment with a digital platform that exogenously changed its rating display rule in a treated neighborhood, which resulted in rating inflation. Using a differences-in-differences approach, we find that platforms benefit from one aspect of rating inflation: user purchases and seller sales increase due to the increased average rating. However, they also face negative consequences: rating inflation causes a decrease in user trial and a greater concentration of sales among popular restaurants. Overall, our results illustrate the potential consequences of rating inflation that platforms need to consider when designing and managing their rating system.

Key words: online ratings, rating inflation, quasi-experiment, differences-in-differences
History:

1. Introduction

Online ratings and reviews are pervasive and influential; more than four out of five U.S. adults consult online ratings and reviews before making purchases (Center 2016). Online ratings reduce the cost of searching for high-quality products and improve product fit (Hong and Pavlou 2014) by allowing consumers to learn from other consumer's past experiences (Duan et al. 2008a, Tadelis 2016). The effectiveness of such social learning depends on the informativeness of online ratings. In this paper, we investigate how a threat to rating informativeness - rating inflation - impacts platforms and their users.

Rating inflation manifests as a combination of high ratings and a low rating variance across sellers. For instance, on eBay, the median seller has a 100% positive rating and the bottom 10^{th}

percentile seller has a 98% positive rating (Nosko and Tadelis 2015). Not only are these exceptionally high positive ratings, but more importantly, there is very little variation between good and mediocre sellers, making these ratings less informative for consumers. One of the earliest works discussing the possibility of inflated ratings was in the Information Systems literature and identified self-selection as a possible mechanism (Li and Hitt 2008). Such rating inflation is common; it has been observed across a variety of digital marketplaces such as online labor markets (Filippas et al. 2018), e-commerce platforms (Nosko and Tadelis 2015) and sharing economy platforms (Zervas et al. 2020). In some settings, rating averages decrease with time (Li and Hitt 2008, Dai et al. 2018, Godes and Silva 2012). But while rating averages may increase or decrease, in this paper, we focus on an important associated phenomenon - the decrease in rating variance across sellers on the platform. Such a decrease in rating variance results in more sellers having identical or similar ratings, making ratings less informative. This can have an impact on the likelihood of users trying a seller for the first time, which we define as *trial*, and which in turn may affect the sales concentration across sellers.

Despite the importance of examining the impact of rating inflation, empirical evidence is limited. A major reason is that quantifying the consequences of rating inflation is challenging. Observational data is not suitable since rating inflation often manifests gradually over time, making it difficult to eliminate all other unobservable temporal confounds, such as improvements in quality, that also influence consumer choices. Recent Information Systems studies have used randomized experiments or natural experiments to investigate the motivations, mechanisms, and impact of user generated content in a variety of settings (Burtch et al. 2018, Huang et al. 2019a,b, Shukla et al. 2021). In our context, a randomized experiment is not feasible since it would pose a risk to the platform's credibility with both sellers and users if ratings were randomly altered. Randomization would also violate the stable unit treatment value assumption (SUTVA) required for identification, since randomly assigning sellers to treatment and control groups would result in the same customer viewing both treated and control sellers affecting each other's outcomes. A natural experiment is also not ideal since they usually do not withhold a control group from being treated which can be used to control for temporal confounds.

In this paper, we conduct a quasi-experiment to empirically examine the impact of rating inflation on user purchases, trial, and sales concentration in the context of choosing restaurants on a food delivery platform. We overcome the identification challenges discussed above by designing and running an experiment with a food delivery platform that induced a rating inflation shock for a treated geographic region while keeping ratings unchanged in other regions. To prevent the treated users from being exposed to both treated and untreated restaurants at the same time, we selected the treated restaurants from a geographically contiguous region that is relatively isolated from the rest of the city.

The experiment and the transaction-level data provide us a unique opportunity to study the consequences of rating inflation on users, restaurants and the platform. We use the differences-in-differences approach to estimate the effect of rating inflation and its impact on user purchases, trial and sales concentration. The geographical selection of the treated group may lead to the presence of systematic differences in the characteristics of treated and control users and restaurants. We validate our results by checking for pre-experiment parallel trends, and running robustness checks using randomization inference and synthetic controls.

Rating inflation is composed of an increase in the rating of the average seller, as well as a decrease in the rating variance across sellers. The increase in average ratings are likely to increase purchases. However, the impact of the decrease in rating variance, can be more nuanced. The decrease in rating variance may increase both the risk, as well as the reward, of trial. The risk increases because rating inflation makes ratings a less informative signal of restaurant quality and increases the uncertainty associated with trial. The reward increases because when ratings are less informative, users rely more on trial to learn about restaurant quality. We empirically investigate whether risk or reward dominates by estimating whether trial decreases or increases using the differences-in-differences method with data generated from the experiment. In addition, due to rating inflation, users rely more on their prior beliefs and experiences in choosing restaurants. They are thus more likely to purchase from more popular restaurants for which they have such signals, resulting in an increase in sales concentration.

We find that although platforms benefit from one aspect of rating inflation - the increase in mean rating - through increased purchases, they may face negative consequences from the other - the decrease in rating variance - through reduced trial and increased sales concentration. The increase in trial risk dominates the increase in trial reward so that consumer trial decreases due to rating inflation. At the same time, rating inflation concentrates sales towards more popular restaurants, increasing such restaurants' market power relative to the platform. Thus, rating inflation makes consumers less willing to try new restaurants and confine themselves to more popular restaurants.

Our findings have important managerial implications on rating system design and platform management. We find that, although rating inflation makes sellers appear of higher quality and may boost total purchases, it can potentially hurt the platform's long-term growth in two ways: it can discourage consumers from trying new sellers, and it may increase the market power of popular restaurants. Overall, our results illustrate the consequences and trade-offs of rating inflation, which can be helpful for platforms when designing and managing their rating system.

2. Related Literature

Our study relates to the stream of Information Systems literature that has examined the factors influencing the creation of user-generated content. Online reviews have been shown to have a role in reducing product fit uncertainty (Hong and Pavlou 2014), and the difference between expectations and actual experience has been shown to affect user's rating decisions (Ho et al. 2017). Similarity of personality traits have been shown to increase the influence of word-of-mouth on consumers (Adamopoulos et al. 2018).

Our work relates methodologically to the stream of Information Systems literature conducting randomized or quasi-experiments or exploiting natural experiments to uncover the mechanisms by which online ratings and review systems influence consumers. Randomized experiments have been used to show that financial incentives increase the volume of reviews given by users while social incentives motivate users to leave longer reviews (Burtch et al. 2018), and that cooperatively framed feedback is most effective at motivating female subjects while competitively framed feedback is effective at motivating male subjects (Huang et al. 2019b). The impact of the implementation of a word-of-mouth system has been investigated through field experiments in an e-commerce setting (Huang et al. 2019a), through a quasi-experiment in a social network setting (Wang et al. 2018), and through natural experiments in healthcare settings (Khurana et al. 2019, Shukla et al. 2021). Our work contributes to this literature by inducing rating inflation through an experiment, on a hyper-local food delivery platform, and presenting its impact on user purchases, trial, and sales concentration.

We contribute to the literature on rating inflation, in which ratings become less informative and useful over time due to a decrease in their variance (Filippas et al. 2018). Most studies on this topic focus on the reasons why the rating averages change over time, some of which are: self-selection (Li and Hitt 2008), reciprocity (Dellarocas and Wood 2008, Bolton et al. 2013, Fradkin et al. 2018, Proserpio et al. 2018) and the related concept of 'reflected' costs (Filippas et al. 2018), herding behavior (Salganik et al. 2006, Muchnik et al. 2013, Aral 2014), and social nudging (Wang et al. 2018). Our study differs from prior studies because instead of investigating the various *causes* of rating average changes, we focus on identifying the *consequences* of rating inflation on user choices. Rating variance has been studied within a product, and has been shown to be correlated with higher demand for low-rated products (Sun 2012), while our focus is on how rating variance across sellers impacts user purchases, trial, and sales concentration.

Our work builds on several recent studies that have focused on improving the design of online rating systems to make them more useful to users. While Chen et al. (2018) show that multi-dimensional rating systems can be more informative to users, Dai et al. (2018) argue that since most users are inattentive, the aggregation of ratings into a single metric is optimal. Similarly, Nosko and Tadelis (2015) demonstrate that adjusting a single metric of seller reputation to make it more informative can improve consumer outcomes significantly. Kokkodis (2019) recommends a rating deflation method to counteract the loss of informativeness due to rating inflation. Our study provides evidence of the need for platform designers to build informativeness into the design of rating systems and estimates the potential consequences if rating inflation were left unaddressed.

More broadly, our work relates to the vast literature that has examined the impact of digital word-of-mouth on sales and found a largely positive effect (Chevalier and Mayzlin 2006, Duan et al. 2008b, Zhu and Zhang 2010, Anderson and Magruder 2012, Lu et al. 2013, Mayzlin et al. 2014, Lewis and Zervas 2016, Tadelis 2016, Song et al. 2019). These studies mostly used observational and often aggregated data to measure the effects of digital word-of-mouth, while we use transaction-level data from an experiment to identify the impact of rating inflation on purchase, trial, and sales concentration. In particular, we contribute to literature that investigates the mechanism of social learning through digital word-of-mouth (Cai et al. 2009, Cabral and Hortacsu 2010, Zhao et al. 2013, Wu et al. 2015, Huang et al. 2016, Acemoglu et al. 2017, Wang et al. 2018, Fang 2022).

3. Data

For this study, we partnered with a large hyper-local food delivery platform in Asia. Our data consists of every transaction made on the chosen food delivery platform in a large Asian city over a period of about 16 weeks. During the observation period, we conducted an experiment in which the rating aggregation rule on the platform was exogenously changed, which resulted in rating inflation. The experiment occurred in a subarea of the city in April 2017, in the 11^{th} week of our observation period. We drop observations for that week to allow for clear demarcation of pre- and post-experiment periods. We also drop observations from the first and last week for which we have incomplete data. All together, we use 9 weeks of pre-experiment data (weeks 2-10) and 4 weeks of post-experiment data (weeks 12-15).

We further refined the dataset in two ways. First, users continued to join the platform during the observation period, but since our approach is to measure changes in behavior due to the experiment, we focused on users who made at least one purchase on the platform one month prior to the experiment. In Appendix B, we analyze new users who join during the observation period, before and after the experiment. Second, we dropped user accounts that have more than three purchases from the same restaurant on the same day. These are likely shared group accounts and as such behave differently than individual users. In Appendix C, we report results from robustness checks with different thresholds. After dropping such users, we were left with 198,044 users who placed 1,510,739 transactions from 2,244 distinct restaurants as summarized in Table 1.

Our main dependent variables for the user-level analysis are weekly purchases ($n\_purchase$) and trials (n_trial). Weekly purchases are a count of the number of purchases made by a user on the

Table 1 Data Characteristics

Description	N
Number of users	198,044
Number of restaurants	2,244
Number of transactions	1,510,739
Number of full pre-experiment weeks	9
Number of full post-experiment weeks	4

platform each week. We measure a user's trial by counting the number of restaurants a user tries for the first time in our observation period each week. We calculate this metric by generating a cumulative count of the number of distinct restaurants a user has tried each week, and denoting the increase in this number as *n_trial* for that week. For example, if a user makes 5 purchases in a week, 3 of which are from restaurants they have previously purchased from, and the remaining 2 are from distinct restaurants they are purchasing from for the first-time within the observation period, then *n_purchase* = 5 and *n_trial* = 2 for that user-week. Note that *n_trial* is defined conditional on purchasing. If *n_purchase* = 0, then *n_trial* is undefined.

Table 2 describes the variables and presents summary statistics for the weekly aggregated data. In our sample, an average user made approximately 0.64 transactions per week. Conditional on making a purchase, the average user trials 0.92 restaurants per week. This relatively high frequency of trial is a consequence of the fact that in our data, we do not observe each user's entire transaction history on the platform; instead, we observe only transactions made within the observation period. To improve the identification of trials, we used the first four weeks of data to generate a history for each user, and used the next five weeks prior to the experiment as our pre-treatment data in our analysis. Doing this improves the accuracy of our count of trial.

The average restaurant has 67.28 transactions per week on the platform, which we denote as their *sales*. Before the experiment, the average restaurant rating on the platform was 3.53.

Table 2 Summary Statistics (per week)

Variable	Description	Mean	St. Dev.	Min	Max
n_purchase	Purchases by user	0.64	1.23	0	29
n_trial	New restaurants tried by user	0.92	1.00	0	14
sales	Number of transactions per restaurant	67.28	99.87	0	1777
pre_rating	Average restaurant rating pre-experiment	3.53	0.28	2.1	4.40

4. Institutional Setting

4.1. App, Ratings, and User Feedback

When registered users open the platform's mobile app or visit their website, they can view restaurants within a 5 km (3.1 miles) radius of their delivery location. Most users, around 85%, transact with the platform through its mobile app while the rest transact through the desktop website. On the home screen of the app or the desktop website, users view a list of available restaurants as well as their rating, estimated time to deliver, and their price range. An illustrative diagram of the mobile app's user interface is shown in Figure 1. Selecting a restaurant displays its full menu from which users select items and complete their order.

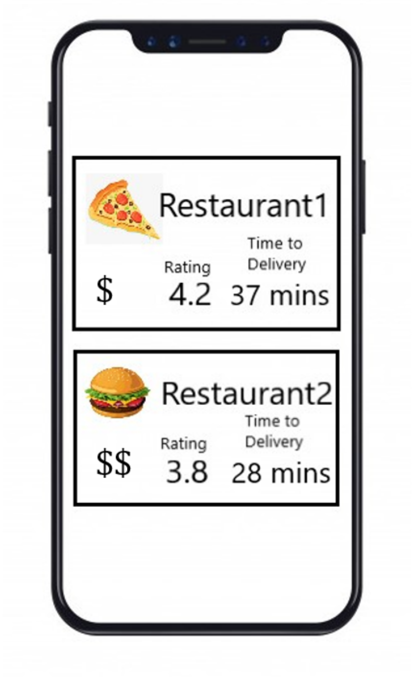


Figure 1 **Illustrative diagram of the delivery platform's mobile app. The restaurant rating is displayed in the bottom center of each listing. To the left of the rating is the price range of the restaurant, and to its right is the estimated time to delivery based on the user's location.**

Ratings are calculated by aggregating user feedback. Users provide a feedback score for their previous transaction on a scale of 0 to 5 stars for restaurant quality before they can initiate a new transaction on the platform. The online platform aggregates these user feedback scores into a numerical rating displayed for each restaurant. Ratings are updated daily to include new user feedback scores received each day.

To allow a restaurant's rating to reflect its current quality, the platform multiplies each user feedback score with a recency-weight. The recency-weight is 1 for feedback received in the most recent 15 days and this weight is reduced by 0.1 for each previous 15-day duration. So, feedback scores received 15 to 30 days ago are weighted by 0.9, those received 30 to 45 days ago are weighted by 0.8, and so on. Feedback scores received over 150 days ago are discarded. This weighted-average rating is then rounded-off to the nearest one decimal place.

4.2. Rating Informativeness and Deflation

The rounding-off of calculated ratings to the nearest one decimal place may lead to a situation where several restaurants have the same displayed numerical rating. For example, all restaurants with calculated ratings between 3.85 and 3.94 are rounded-off to have a displayed rating of 3.9. To appreciate the severity of this issue, consider that out of 48 restaurants in one neighborhood, 13 had the same rating of 4.3, another 13 had a rating of 4.2, and 9 had a rating of 4.1. Together, almost three-fourths of all restaurants in the neighborhood had ratings in the narrow range of 4.1 and 4.3. For users residing in this neighborhood, ratings provided limited information to differentiate between available restaurants. Due to their high mean and low variance, we refer to these as "inflated" ratings.

With this context, we consider rating informativeness to be determined by the variance of the rating distribution. A rating system is informative to the extent that it helps users differentiate between restaurants; distinct ratings are more informative than identical ratings. The greater the rating variance, the lower the number of restaurants with identical ratings, making ratings more informative for users. In Appendix A, we provide a brief overview of commonly used rating systems on other platforms and their respective informativeness.

The platform, about 14 months before the observation period of this study, recognized the problem of less informative ratings, and sought to ameliorate it by increasing the variance of the rating distribution. This change reduced the number of restaurants having identical ratings. At the same time, to accommodate the higher rating variance after redistribution, the platform also decreased the mean of the rating distribution. Thus, ratings were artificially *deflated* by the platform to make them more informative 14 months before the observation period of this study. This approach of adjusting ratings to make them more informative is recommended by Dai et al. (2018), Nosko and Tadelis (2015), and Kokkodis (2019).

4.3. Experiment: Rating Inflation

Rating deflation by the platform about 14 months before the observation period of this study was an attempt to counteract rating inflation. Note that 14 months is a relatively long time so we do not expect that the rating deflation shock has an impact on our study. We utilized this unique opportunity to exogenously induce a rating inflation shock for this study. We dropped the first week's data as it was incomplete. After 9 weeks of pre-experiment data, we rolled-back the rating deflation imposed by the platform for a treated neighborhood in week 11, while retaining the artificially deflated ratings for the rest of the city, as a control group. Thus, the treated group experienced a rating inflation shock while the control group did not. Rolling-back the rating deflation caused a temporary platform-wide disruption to the service for two days that decreased orders on the platform from all restaurants. This disruption was quickly fixed and order volume reached normal levels within two days. Since the disruption was platform-wide, it does not have an impact on our identification strategy. We drop the data from week 11 for our analysis to conservatively exclude the period around the disruption. We observe post-experiment data for four full weeks after the experiment. The timeline of rating changes relative to the observation period is illustrated in Figure 2.

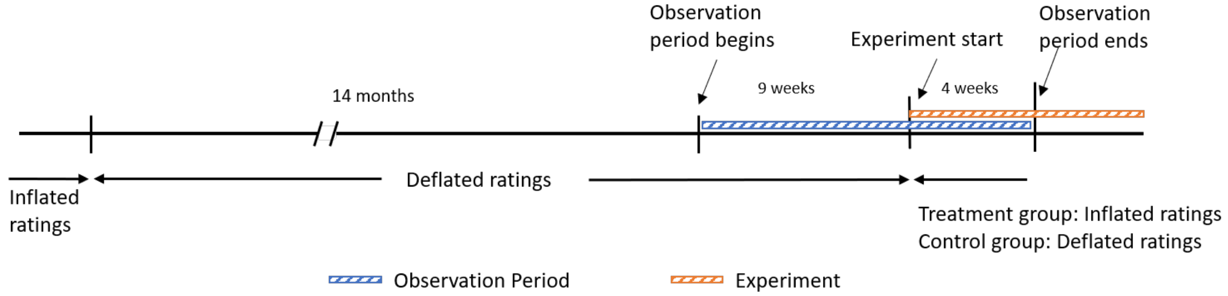


Figure 2 Timeline for rating changes. Original inflated ratings were changed to deflated ratings 14 months before the observation period. First week's data is dropped for being incomplete. Ratings were inflated for the treated region in week 11 of the observation period. During the experiment, inflated ratings were displayed for the Treated neighborhood while deflated ratings continued to be displayed for Control neighborhoods.

The treated neighborhood had 48 restaurants on the platform from which 3,753 distinct users made purchases in the pre-experiment observation period. To prevent the treated users from being exposed to both treated and untreated restaurants at the same time, we selected the treated restaurants from a geographically contiguous region that is relatively isolated from the rest of the city. While such geographical selection of the treated group may lead to the presence of systematic differences in the characteristics of treated and control users and restaurants, we validate our results by checking for pre-experiment parallel trends, and running robustness checks with randomization inference and synthetic controls. More details are discussed in Sections 6.2.3 and 7.4.

For treated group restaurants, the experiment increased the average rating by 0.71 stars on a 5-star rating scale going from 3.45 stars to 4.16 stars. At the same time, the variance of the rating distribution decreased by 55% , dropping from 0.060 to 0.027. Together, these changes can be viewed as rating inflation. The decrease in rating variance is reflected in the reduced range of ratings for the 48 treated restaurants: the ratings take 12 distinct values before the treatment and only 8 distinct values after. Therefore, more restaurants have identical ratings after the rating inflation, making ratings less informative. The rating distributions are shown in Figure 3.

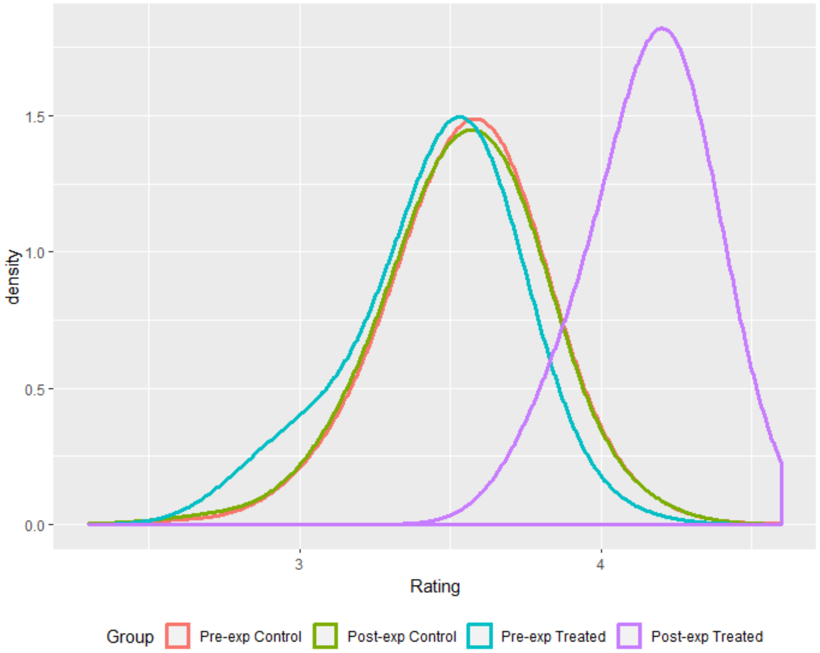


Figure 3 **Rating distribution before and after the experiment for Treated and Control groups. Treated group experiences rating inflation while the Control group does not.**

Since, as discussed earlier, treatment cannot be randomized, there are some differences in the activity levels of users in the treated and control groups. In Figure 4, we show how the proportion of users at various levels of purchase and trial varied over the observation period. We see that the trends are largely similar for purchases, while the experiment has a discernible impact on trials for treated users after the experiment. We provide a more formal check of parallel pre-experiment trends in Section 6.2.3. Besides user activity levels, we further compare neighborhood totals (total weekly purchases and trials) of the treated and control groups in Appendix D.

5. Hypotheses Development

The rating inflation shock induced by the experiment can be decomposed into an increase in the average rating and a decrease in the variance of ratings for treated restaurants. Here we describe how these two changes affect treated users' perception of restaurant quality, and consequently, their choices.

The increase in average rating due to rating inflation results in treated users perceiving restaurants on the platform to be of higher average quality than before. The decrease in rating variance reduces the informativeness of rating signals which increases the uncertainty of a user's perception of restaurant quality. Since the rating signal is less certain, it becomes less important in shaping a user's quality perception. Correspondingly, the importance of other signals, such as prior beliefs and prior usage experience increase. Therefore, due to rating inflation, users rely less on ratings, and

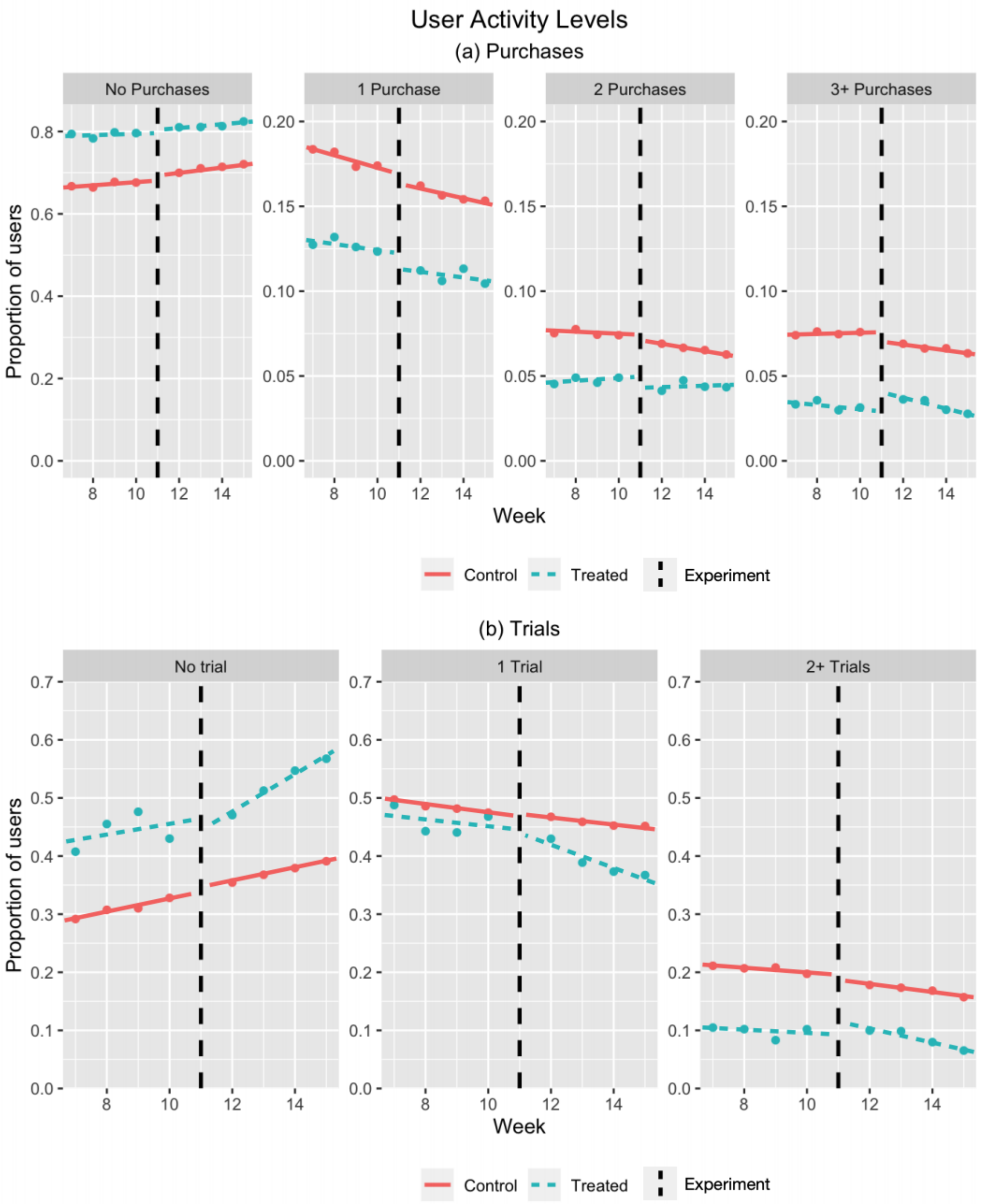


Figure 4 **User Activity Levels. The experiment occurs in week 11 and pre-experiment trends are parallel for treatment and control group users.**

(a) Top row shows the proportion of users who make no purchase, 1, 2, or 3 or more purchases each week.

(b) Bottom row shows the proportion of users who have no trial, 1 trial, or 2 or more trials each week. For the treated group, the proportion of users with no trial increases and the proportion of users with 1 or more trials decreases after the experiment, indicating that trials decrease on average after the experiment.

more on their prior beliefs and usage experiences. In this section, we connect the changes in user quality perception due to rating inflation with the expected observable effects on user purchases and trial, and correspondingly, on restaurant sales and sales concentration.

5.1. Effect on User Purchases and Restaurant Sales

Our study examines the effect of rating inflation on experiential service of restaurant and food delivery. Other studies have examined similar effects in online retail. E-commerce websites saw a higher likelihood of products being added to carts when a larger number of Word-Of-Mouth comments were seen by buyers (Huang et al. 2019a). Books on Amazon.com and Barnesandnoble.com sold more copies when ratings were raised (Chevalier and Mayzlin 2006).

In the context of restaurants, studies have found that restaurant sales increase merely by being listed on review platforms (Lu et al. 2013) and this value can be quantified (Wu et al. 2015). Further, higher ratings have been shown to increase sales in a variety of settings. On Yelp.com, restaurants with a half-star higher rating sell out 19 percentage points more often, and the effect is even greater when customers only have information from this platform (Anderson and Magruder 2012).

In our context, the increase in mean rating due to rating inflation increases the perceived quality of restaurants on the platform. Assuming the perceived utility of the outside option does not change, we expect users to purchase more often on the platform after rating inflation and the restaurant sales to increase. The decrease in rating variance may make it harder for users to choose between restaurants and, thus, may shift their consumption patterns across restaurants, but overall, we expect that the increase in mean rating boosts the overall sales on the platform so that rating inflation increases user purchases and restaurant sales.

H1: *Rating inflation increases user purchases and restaurant sales.*

5.2. Effect on User Trial

Users rely more on ratings when trying a new restaurant because of the absence of any direct experience with the restaurant. Inflated ratings can be deflated to create more accurate estimates of a seller's quality, allowing customers to make better judgements (Kokkodis 2019). While literature has made it clear that inflated ratings reduce informativeness of ratings, we further conjecture that this results in fewer user trials. In particular, rating inflation can affect both the risk of trial and the reward from trial, as we discuss below.

Increase in Risk of Trial: With rating inflation, rating variance decreases and rating signals become less informative. Users find it more difficult to distinguish between restaurants and are less certain about restaurant quality (Dai et al. 2018, Nosko and Tadelis 2015). This is especially significant for restaurants they have not tried previously. Reviews have been shown to increase the sales of high-quality independent restaurants and facilitate user learning about restaurant quality, especially for tourists and travelers who are more likely to engage in trial than locals (Fang 2022). Similarly, rating inflation, by increasing the uncertainty of restaurant quality, increases the risk of trial.

Increase in Reward from Trial: At the same time, the decrease in rating variance leads to an increase in reward from trial. We consider the reward from trial to be the value of the information gained by trial . In general, consumers can gain information about new restaurants either through the ratings and reviews or by trying the restaurants themselves. When rating inflation occurs, the value of the information from the rating system decreases, so the value of the information from trial becomes relatively more valuable (i.e., the reward of trial becomes larger) (Acemoglu et al. 2017). To see how this works, consider an extreme case of rating inflation, where all restaurants have the same rating on a platform. Now, to make better choices, users must engage in trial to gather usage signals about restaurant quality. This gathering of usage signals has become more important, and hence more valuable, when ratings are uninformative. Therefore, trial becomes more valuable due to rating inflation.

The increase in reward from trial can be explained by the cue diagnosticity theory (Feldman and Lynch 1988, Dimoka et al. 2012): the degree to which consumers rely on and use a specific cue in a decision depend on the cue's diagnosticity. If a cue is nondiagnostic, consumers will turn to alternative cues that they find to be diagnostic. When rating inflation happens, rating becomes a less diagnostic cue, so consumers will engage in trial to gain own experience as an alternative cue (Yi et al. 2017) or may even limit their adoption (Cenfetelli and Schwarz 2011).

Overall, both the risk of trial and the reward from trial increase due to rating inflation. Which of these two effects dominates, i.e. whether trial decreases or increases, is tested empirically through the experiment. We expect that the increase in risk from trials dominates the increase in reward from trials, so consumers will reduce trials after the experiment.

H2: *Rating inflation decreases user trial.*

5.3. Effect on Restaurant Sales Concentration

The changes in user choices can have an impact on the distribution of sales across restaurants on the platform. Understanding how rating inflation affects sales concentration is important for the platform because if sales concentrate among the top few restaurants on the platform due to rating inflation, the platform's market power relative to these top restaurants may decrease. Such restaurants may be able to negotiate a lower commission to be paid to the platform for each transaction. On the other hand, greater sales concentration may lower operational costs for the platform by combining multiple orders for delivery. In either case, it is important for the platform to understand the factors affecting sales concentration.

Rating inflation, by making rating signals less informative, makes users rely more on substitute signals, namely their prior beliefs and usage experience. Assuming offline and online restaurant popularity are correlated, popular restaurants are more likely to have been heard of (i.e. prior beliefs) or experienced by users (i.e. usage experience) online or offline. Therefore, users are more likely to have informative prior beliefs and usage experience signals for more popular restaurants, and as a result, their quality perception would be more precise for popular restaurants. As such, in response to rating inflation, users are likely to shift their consumption to more popular restaurants, and we expect the sales concentration to increase.

The differential impact of rating system on heterogeneous sellers has recently been examined by other studies. It was found that by giving consumers access to a rating system such as that of Yelp.com, high quality restaurants saw an increase in sales while low quality restaurants experienced a decrease (Fang 2022). Hotels with higher ratings on Yelp or Trip Advisor had higher demand and were able to charge higher prices (Lewis and Zervas 2016). In another study on a doctor appointment booking platform, it was found that doctors who were highly rated benefited at the expense of unrated doctors (Shukla et al. 2021). These studies focus on how rating system affects different sellers and find that popular or higher-rated sellers benefit more. Our study focuses on how a change in the rating system (i.e., rating inflation) affects different sellers. We expect a similar effect: rating inflation benefits popular restaurants the most and increases sales concentration on the platform.

H3: *Rating inflation increases restaurant sales concentration.*

6. Methodology

6.1. Empirical Model

6.1.1. For Users For platform users, the dependent variables we are interested in analyzing - *n_purchases* and *n_trial* - are count variables. As such, we conduct our analysis using the pseudo-maximum likelihood fixed-effects Poisson regression model (Ciani and Fisher 2018, Silva and Tenreyro 2006, 2011) for its several desirable properties: suitability for non-negative but skewed data (Azoulay et al. 2010), consistency (Wooldridge 2010), and robustness to arbitrary patterns of serial correlation (Wooldridge 1997). The last property allows us to use weekly aggregated data rather than aggregating at the pre/post-experiment level (Bertrand et al. 2004). This helps in further controlling for temporal trends in our data (Wang and Goldfarb 2017). An additional benefit of using the Poisson estimator is that it does not suffer from the incidental parameters problem (Wooldridge 2010, Cameron and Trivedi 2013, Fernández-Val and Martin 2016). We estimate the coefficients using the PPMLHDFE command in Stata (Correia et al. 2020), which allows for fast estimation of pseudo Poisson regression models with high-dimensional fixed effects. It is robust to statistical separation (Correia et al. 2019) and singletons (Correia 2015) which are present in our data.

We estimate the following equation to get the average treatment effect of the rating inflation experiment:

$$Y_{ist} \sim \text{Poisson}[\mu_i \, exp(\beta D_{st} + \tau_t)] \tag{1}$$

where $Y_{ist} \in \{n\_purchase_{ist}, n\_trial_{ist}\}$ is the dependent variable for user i, group $s \in \{treated, control\}$, and week t. μ_i is the user fixed effect and τ_t captures the week fixed effects. D_{st} is the treatment indicator equal to one for treated groups after the experiment, and zero otherwise and β is the coefficient of interest. Cognizant of the challenges in interpreting interaction term coefficients in non-linear models (Ai and Norton 2003, Puhani 2012), we interpret this result as a difference-in-semielasticity (DIS), defined as "the second explanatory variable's impact on the dependent variable with respect to the first explanatory variable" (Shang et al. 2018).

6.1.2. For Restaurants The dependent variable of interest when analyzing restaurants on the platform is the number of weekly transactions - *sales*. We identify the average treatment effect on sales with the following equation:

$$sales_{jst} = \alpha + \beta D_{st} + \lambda_j + \tau_t + \epsilon_{jst} \tag{2}$$

where $sales_{jst}$ is the number of transactions on the platform for restaurant j, belonging to group $s \in \{treated, control\}$, in week t. α is a constant and β is the estimate of the average treatment effect. D_{st} is the treatment indicator equal to 1 for treated restaurants after the experiment and 0 otherwise. λ_j and τ_t are the restaurant and week fixed effects, and ϵ_{jst} is the error term.

Next, to identify the change in sales concentration on the platform, we estimate the heterogeneous treatment effect of the rating inflation experiment on restaurants according to their popularity using the following equation:

$$sales_{jst} = \alpha_0 + \alpha_1 after_t \, \text{x} \, pre\_sales_j + \beta_1 D_{st} + \beta_2 D_{st} \, \text{x} \, pre\_sales_j + \lambda_j + \tau_t + \epsilon_{jst} \tag{3}$$

where $sales_{jst}$ is the number of transactions on the platform for restaurant j, belonging to group $s \in \{treated, control\}$, in week t. $after$ is a binary variable that equals 1 after the experiment and 0 prior to it. pre_sales_j is the total pre-experiment sales for restaurant j. D_{st} is the treatment indicator that equals 1 for treated groups after the experiment, and 0 otherwise. β_2 is the coefficient of interest; a positive value implies that more popular restaurants experience a larger increase in sales, suggesting an increase in sales concentration on the platform. λ_j and τ_t are the restaurant and week fixed effects, and ϵ_{jst} is the error term.

6.2. Identification Strategy

Two key assumptions are required for credible identification using diferences-in-differences: (i) stable unit treatment value assumption (SUTVA) and (ii) parallel trends. In this section, we discuss our approach to minimize violations of SUTVA and check for pre-experiment parallel trends of our dependent variables. But first, we describe how we filter our data sample to improve the accuracy of our measure of trial.

6.2.1. Data Sample Our data span 9 pre-experiment weeks and 4 post-experiment weeks. Due to this limited observation period, we do not observe each user's entire history of purchases on the platform. We consider the first time a user purchases from a restaurant in our observation period as a "trial," even if they have purchased from that restaurant before the observation period began. This could lead to an inflation in the trial count. For instance, the first purchase of every user in our observation period is always considered a trial in our dataset.

We minimize potential identification problems from this issue in two ways: First, we perform our main analysis on a sample that excludes the initial four weeks of data (weeks 2 to 5), which are mostly likely to over-count trial. Second, we emphasize that our estimation strategy does not require an accurate count of trial, but instead relies on changes in trends of trial between treated and control groups. We have the same issue of over-counted trials for both treated and control groups, so we do not expect this issue to affect our diff-in-diff estimate. We validate this assumption by checking whether the trends for trial for treated and control groups were parallel in the pre-experiment period in Section 6.2.3. These steps alleviate the concern for potential bias in our estimates.

6.2.2. Stable Unit Treatment Value Assumption Our experiment was designed to minimize potential violations of SUTVA by choosing a relatively isolated contiguous geographic area for treatment so that few users view restaurants from both the treated and control groups. If a user is shown restaurants from both the treated and control groups, they have an intermediate level of treatment, which is a violation of SUTVA. Given that our treatment is imposed at the restaurant-level and not users, we expect some users will see restaurants from both the treated and control groups.

Seeing restaurants with ratings calculated by two different methods might cause users to have unpredictable reactions. Some may shift their purchase towards the treated restaurants, while others might question the reliability of the ratings on the platform, especially if it contradicts their prior knowledge about the quality of the restaurants. For example, if a restaurant renowned for its quality is displayed with a low rating because of the method by which it has been calculated, while a mediocre restaurant is displayed with a high rating, users might disregard the rating entirely, or may even distrust the platform. We attempted to mitigate this risk by selecting the treated restaurants from a geographically contiguous region that is relatively isolated from the rest of the city. With such a strategy, we attempted to minimize the number of users for whom the 5 km radius contained restaurants from both treated and control groups. While we could not observe the list of restaurants each customer had available to them, we observed their transactions. We found that only 280 users, or about 0.14% of the total, transacted with restaurants from both the treated and control groups. This suggests that the strategy for isolating the treated users from control group ratings was largely successful. We dropped these users from our analysis.

6.2.3. Parallel Trends Assumption While implementing the treatment condition on a geographic basis minimizes potential SUTVA violations, it could lead to violations of the parallel trend assumption. Different geographic regions may have different kinds of users and restaurants that reflect that region's socio-economic and demographic factors. To rule out this possibility, we checked whether the pre-experiment trends for users and restaurants in the treated and control groups were parallel for our dependent variables of interest.

Recall that we dropped the first four weeks of data to allow for a more accurate measure of trial. We also dropped the week of the rating change, which is the 11^{th} week in our data, to get clear pre- and post-treatment periods. We are left with five weeks of pre-treatment observations (weeks 6-10) and four weeks of post-treatment (weeks 12-15). We estimate the lead and lag coefficients of the treatment effect by estimating the following equation:

$$Y_{ist} \sim \text{Poisson}\left[\mu_i \exp\left(\sum_{t=6,\dots,10}^{12,\dots,15} \beta_t week_t \, \text{x} \, treated_s\right)\right] \tag{4}$$

where Y_{ist} is the dependent variable for user i, $t = \{6, ..., 10\} \cup \{12, ..., 15\}$ are weeks, and $s \in \{treated, control\}$ is the group. μ_i are the user fixed effects, $week_t$ is the indicator variable for week t and $treated_s$ is the indicator variable for the treated group. β_t with $t \in \{6, ..., 10\}$ are the 'lead' estimates of the treatment effect. β_t with $t \in \{12, ..., 15\}$ are the 'lag' estimates of the treatment effect.

We plot the estimated coefficients for user purchase ($n\_purchase$) and user trial (n_trial) in Figure 5(a) and find that almost all lead coefficients are statistically insignificant. Thus, we conclude that trends are parallel for the treated and control users prior to the experiment.

To check for parallel trends for restaurant sales, we estimate the lead and lag coefficients of the average treatment effect (β_t) on sales in the following equation:

$$sales_{jst} = \alpha + \sum_{t=6,\dots,10}^{12,\dots,15} \beta_t week_t \, \text{x} \, treated_s + \lambda_j + \epsilon_{jst} \tag{5}$$

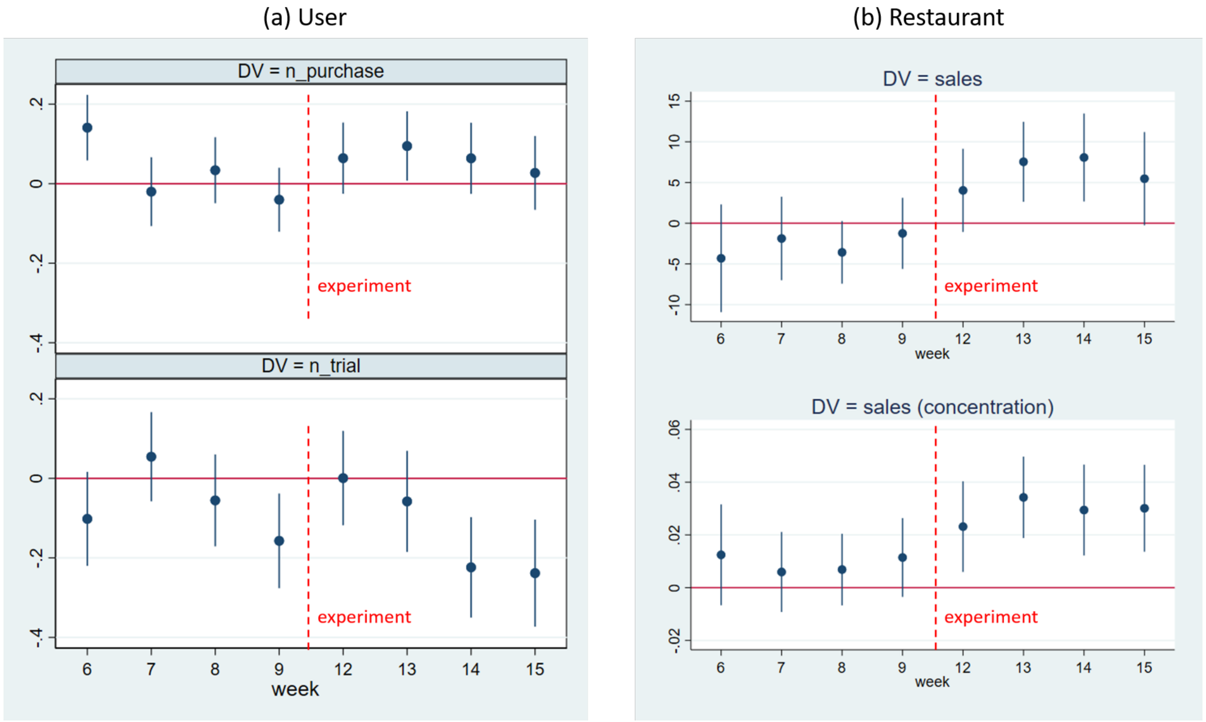


Figure 5 **(a) User purchases and trial: Coefficient plot of Equation (4) for β_t, $t \in \{6,...,9\} \cup \{12,...,15\}$.**
(b) Restaurant Sales: Coefficient plot of Equation (5) for β_t, $t \in \{6,...,9\} \cup \{12,...,15\}$. Restaurant Sales Concentration: Coefficient plot of Equation (6) for γ_t, $t \in \{6,...,9\} \cup \{12,...,15\}$.
Week $t = 10$ is the baseline and its coefficient is normalized to zero. Displays the 95% confidence intervals.

where $sales_{jst}$ is the number of transactions on the platform for restaurant j, belonging to group $s \in \{treated, control\}$, in week t. $week_t$ is the indicator variable for week t and $treated_s$ is the indicator variable for the treated group. β_t with $t \in \{6,...,10\}$ are the 'lead' estimates of the average treatment effect. β_t with $t \in \{12,...,15\}$ are the 'lag' estimates of the average treatment effect.

Similarly, for checking parallel trends for restaurant sales concentration, we estimate the lead and lag coefficients of the heterogeneous treatment effect (γ_t) in the following equation:

$$\begin{aligned} sales_{jst} = \alpha_0 + \alpha_1\, after_t \,\mathrm{x}\, pre\_sales_j + \sum_{t=6,...,10}^{12,...,15} \beta_t\, week_t \,\mathrm{x}\, treated_s \\ + \sum_{t=6,...,10}^{12,...,15} \gamma_t\, week_t \,\mathrm{x}\, treated_s \,\mathrm{x}\, pre\_sales_j + \lambda_j + \epsilon_{jst} \end{aligned} \tag{6}$$

where pre_sales_j is the total pre-experiment sales for restaurant j. γ_t with $t \in \{6,...,10\}$ are the 'lead' estimates of the heterogeneous treatment effect. γ_t with $t \in \{12,...,15\}$ are the 'lag' estimates of the heterogeneous treatment effect.

We plot the coefficients β_t in Equation (5) and γ_t in Equation (6) in Figure 5(b) and conclude that the pre-experiment trends for both restaurant sales and sales concentration are parallel for treated and control restaurants.

To summarize, we find that pre-experiment trends for the dependent variables - user purchases, trial, restaurant sales, and restaurant sales concentration – are parallel for treated and control units. We assume that these trends would have remained parallel in the absence of the rating inflation treatment and we can use the control units to estimate the counterfactual for the treated units in the absence of rating inflation. This allows us to have a causal interpretation of the results in the next section.

7. Results

7.1. User Purchases and Restaurant Sales Increase

We expect that users would purchase more often on the platform when faced with inflated ratings. We first show some model-free evidence of user purchases in Figure 6(a). We observe a small relative increase in the average purchases per week for treated users compared with control users after the experiment[1].

We estimate Equation 1 using *n_purchases* as the dependent variable and present the DD estimates in Table 3. Columns (1) and (2) show the results with either week fixed effects or user fixed effects; Column (3) show the results with both week and user fixed effects as in Equation 1. We find that user purchases increase by approximately 3.5% due to rating inflation. While the direction of this effect is as expected, we note that this is a short-term effect in response to a sudden rating inflation. Two caveats are worth noting. First, if rating inflation were to manifest more gradually, as is usually the case, we cannot conclude from these results that a similar increase in user purchases would happen. Second, over time, this increase may taper as users recalibrate their expectations of what high ratings signify on the platform. Nevertheless, in the setting of this experiment, we find evidence that rating inflation causes an increase in user purchases in the short term.

Due to rating inflation, the restaurants on the platform appear more attractive than the outside option, so we expect that restaurant sales increase after the experiment. In Table 4, columns (1) and (2) show the DD estimates of Equation 2, without and with week fixed effects. We find that restaurant weekly sales increase due to the experiment. This is consistent with our previous finding that user purchases increase after the experiment.

[1] While we are unable to definitively determine the source of the declining trend of the control group in Figure 6(a), we conjecture that this could potentially be due to the technical disruption at the roll-out of the experiment that affected both the treatment and control groups. As such, the control group still acts as the counterfactual for the treatment groups' trend.

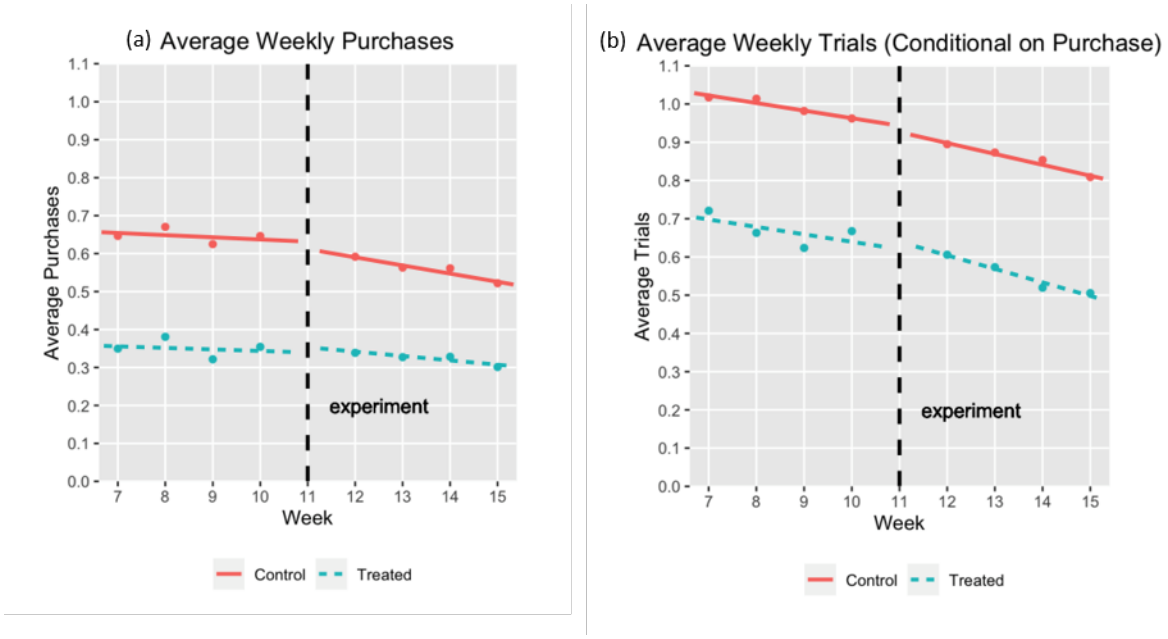


Figure 6 **(a) User Average Weekly Purchases by Group. (b) User Average Weekly Trial by Group. Pre-experiment trends are parallel for treated and control users. Compared to the decline in the control users post-experiment, the treated user purchases do not decline as much, whereas treated user trials decline more relative to the control users.**

Table 3 **User Purchases Increase Due to Rating Inflation**

DV = $n\_purchase$	(1)	(2)	(3)
D_{st}	0.0345	0.0345*	0.0345*
	(0.0283)	(0.0204)	(0.0204)
Week Fixed Effects	Yes	-	Yes
User Fixed Effects	-	Yes	Yes
Observations	1782396	1351863	1351863

Cluster-robust standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4 **Restaurant Sales Increase Due to Rating Inflation**

	All Customers		Only Repeat Customers	
DV = $sales$	(1)	(2)	(3)	(4)
D_{st}	9.937***	9.937***	11.81***	11.81***
	(1.546)	(1.540)	(1.355)	(1.220)
Week Fixed Effects	-	Yes	-	Yes
Restaurant Fixed Effects	Yes	Yes	Yes	Yes
Observations	29198	29198	27183	27183

Cluster-robust standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

7.2. User Trial Decreases

Users face an increased risk of trying a new restaurant as a result of rating inflation. At the same time, users face an increased reward from trying a new restaurant as rating inflation makes information from trial more valuable. We expect that the increase in risk dominates the increase in reward so that user trials decrease after the experiment. The model-free evidence in Figure 6(b) supports this conjecture: there is a relatively steeper decrease in the average trial per week for treated users compared with control users after the experiment.

We estimate Equation 1 using *n_trial* as the dependent variable and present the DD estimates in Table 5. Columns (1) and (2) show the results with either week fixed effects or user fixed effects; Column (3) show the results with both week and user fixed effects as in Equation 1. We find that users reduced their trial in response to rating inflation. This implies that the increase in risk of trial outweighs the increase in reward from trial. The DD estimates in Column (3) show that the number of new restaurants an average user tries decreased by approximately 7% due to rating inflation.

Table 5 User Trial Decreases due to Rating Inflation

DV = n_trial	(1)	(2)	(3)
D_{st}	0.00245	-0.0664**	-0.0670**
	(0.0289)	(0.0301)	(0.0301)
Week Fixed Effects	Yes	-	Yes
User Fixed Effects	-	Yes	Yes
Observations	564634	513136	513136

Cluster-robust standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Given that user trials decrease after the experiment, we expect that restaurant sales come more from repeat customers rather than from new customers. To see whether this is the case, we re-estimate Equation 2 using sales from only repeat customers as the dependent variable and present the DD estimates in columns (3) and (4) of Table 4. We find that restaurant weekly sales from repeat customers increase due to the experiment. The result supports our previous finding that user trials decrease after the experiment.

Note that the decrease in user trial happened in conjunction with the increase in purchases by users. Thus, while rating inflation induced users to purchase more often, they were reluctant to try new restaurants due to less informative ratings. To the extent that the platform seeks to enable users to try new restaurants and discover new favorites, rating inflation inhibits that goal.

7.3. Popular Restaurants Benefit

The decrease in rating informativeness due to rating inflation is expected to result in increase in sales concentration. Users have fewer other signals for less popular restaurants and, as such, might be expected to prefer purchasing from better known and popular restaurants. We provide model-free evidence of this effect in Figure 7. Figure 7(a) presents the distribution of restaurant sales for the control and treated groups before and after the experiment. We find that while the density of the average weekly sales does not shift for the control group after the experiment, it shifts to more popular restaurants for the treated group. Figure 7(b) presents how Herfindahl-Hirschmann Index (HHI), a measure of market concentration, changes over time for the control and treated neighborhoods. We find that HHI increases for the treated neighborhood after the experiment, again suggesting that restaurant sales concentrate more on the popular restaurants for the treated neighborhood after the experiment.

Table 6 Sales of More Popular Restaurants Increase due to Rating Inflation

DV = *sales*	(1)	(2)	(3)
$after$	0	3.018***	0
	(.)	(0.645)	(.)
$treated$	0.226	0	0
	(0.886)	(.)	(.)
pre_sales	0.101***	0	0
	(0.000447)	(.)	(.)
$after$ x pre_sales	-0.00438***	-0.00438***	-0.00438***
	(0.00124)	(0.00109)	(0.00109)
$treated$ x pre_sales	0.0000136	0	0
	(0.00230)	(.)	(.)
D_{st}	-0.438	-0.438	-0.438
	(1.952)	(1.778)	(1.767)
D_{st} x pre_sales	0.0216***	0.0216***	0.0216***
	(0.00470)	(0.00440)	(0.00436)
$constant$	1.167***	67.18***	68.10***
	(0.279)	(0.155)	(0.251)
Week Fixed Effects	Yes	-	Yes
Restaurant Fixed Effects	-	Yes	Yes
Observations	29198	29198	29198

Cluster-robust standard errors in parentheses

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

We estimate Equation 3 using restaurant sales as the dependent variable and present the coefficient estimates in Table 6. *pre_sales* is the total sales of a restaurant prior to the experiment and captures restaurant popularity. The negative coefficients of D_{st} show that rating inflation reduces the sales of less popular restaurants, while the positive coefficients of D_{st} xpre_sales implies that sales increase for more popular restaurants. Taken together, the results suggest a shift of sales from less popular restaurants to more popular restaurants, increasing sales concentration.

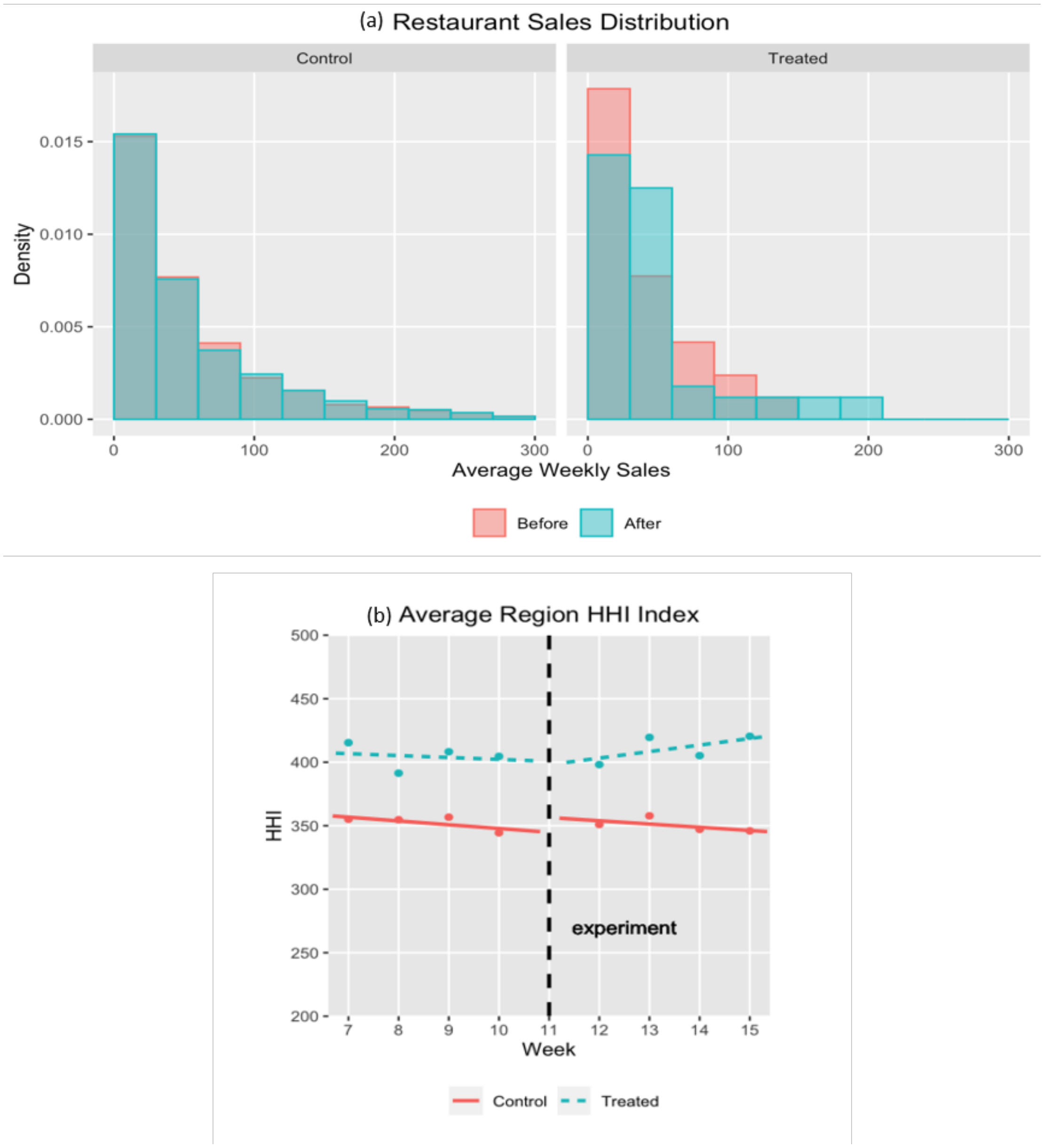


Figure 7 **(a) Sales distribution across restaurants before and after the experiment for the Control and Treated groups. The sales distribution does not change for the Control group whereas sales shift to more popular restaurants in the Treated group.**

(b) Herfindahl-Hirschmann Index (HHI) for treated and similar control neighborhoods. Pre-experiment trends for sales concentration are parallel. Sales concentration increases for the treated neighborhood after the experiment.

We interpret the results in Table 6 as a difference-in-semielasticity (DIS) as $\exp(-0.438 + 0.0216) + \exp(-0.438) = 1.4\%$ as in Shang et al. (2018). This implies that the experiment increased sales by 1.4% for each additional unit of total pre-experiment sales.

The increase in market concentration can be explained by two observations. As rating inflation makes ratings less informative, users rely more on other signals, in particular their prior experience, which explains why sales shift towards more popular restaurants. In addition, we find that trial decreases due to rating inflation, so users make repeat purchases from the most popular restaurants.

7.4. Robustness Checks

Our main results so far have relied on the parallel-trend assumption, evidence of which we find by checking for pre-experiment trends in Section 6.2.3. However, we note that because the treatment and control regions are geographically separated regions in a large city, it is possible that there is a difference in the quality and quantity of restaurant options available to users in control and treatment regions. This fact, by itself, does not pose a challenge for identification as long as the parallel trends assumption holds. We, nevertheless, perform two robustness checks - (i) randomization inference, and (ii) synthetic controls below. A third robustness check as a pre-experiment placebo test is reported in Appendix C2.

7.4.1. Randomization Inference We check the robustness of our estimates by calculating empirical p-values using randomization inference (Imbens and Rubin 2015). For each model, we generate a randomized treatment vector and estimate the regression equation. We perform the randomization 1,000 times for each model and plot the generated coefficient estimates with the actual estimate of the coefficient in Figure 8 . The empirical p-value is the proportion of generated coefficients with as or more extreme values than the actual coefficient. Table 7 shows the results of the randomization inference, with mean values of the randomization estimates, its standard deviation, and the empirical p-value. The empirical p-values are evidence that our estimates are robust.

Table 7 Results from Randomization Inference with Empirical P-values

	Users		Restaurants	
	(a) Purchase	(b) Trial	(c) Sales	(d) Sales_conc
Mean of Random β	0.00119	0.000721	-0.1225	-0.000415
Std Dev of Random β	0.0215	0.0219	3.914	0.00914
Replications	1000	1000	1000	1000
Estimated β	0.0345	-0.0670	9.937	0.0216
Empirical P-Value	0.054	0.001	0.017	0.005

7.4.2. Synthetic Controls We perform another robustness check of our results by creating synthetic control units to estimate the treatment effect. Our treatment is localized to one neighborhood. We use neighborhood- and user-level characteristics to create synthetic control users that are similar to the treated users. While the user-level characteristics match user behavior,

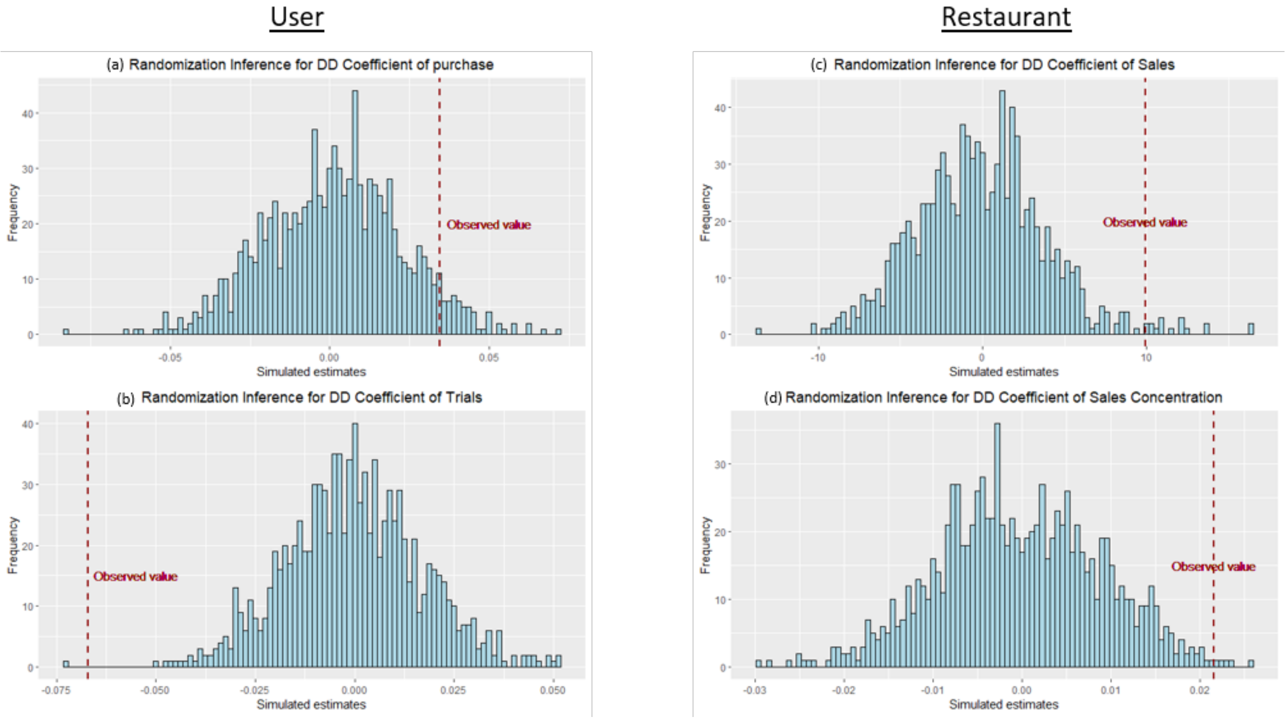


Figure 8 **Randomization inference for (a) user purchases, (b) trials, (c) restaurant sales, and (d) sales concentration. The empirical p-values are 0.054 for purchases, 0.001 for trials, 0.017 for sales, and 0.005 for sales concentration.**

the neighborhood-level characteristics match the environment in which the users find themselves in. For the neighborhood-level characteristics, we calculate the average daily sales and ratings of restaurants in the neighborhood and the average transaction amount. These variables capture the average popularity, quality, and price-tier of restaurants in the neighborhood that are available to the user. For the user-level characteristics, we include the average amount spent by the user prior to the experiment to create synthetic units that spend similarly on the platform. We also include the weekly values of the variables of interest, *n_purchase* and *n_trials*, for the pre-experiment period. The results are shown in Figure 9 and Table 8. We find that purchases increase and trials decrease after the experiment when using synthetic control units to run the analysis.

Together, the randomization inference and synthetic control results add further confirmation to the robustness of our results.

Table 8 **Results from synthetic control user analysis**

	Percent Change	p-value	Lower Bound	Upper Bound
n_purchase	5.9%	0.434	-6.0%	19.2%
n_trial	-24.9%	0.039	-40.3%	-5.6%

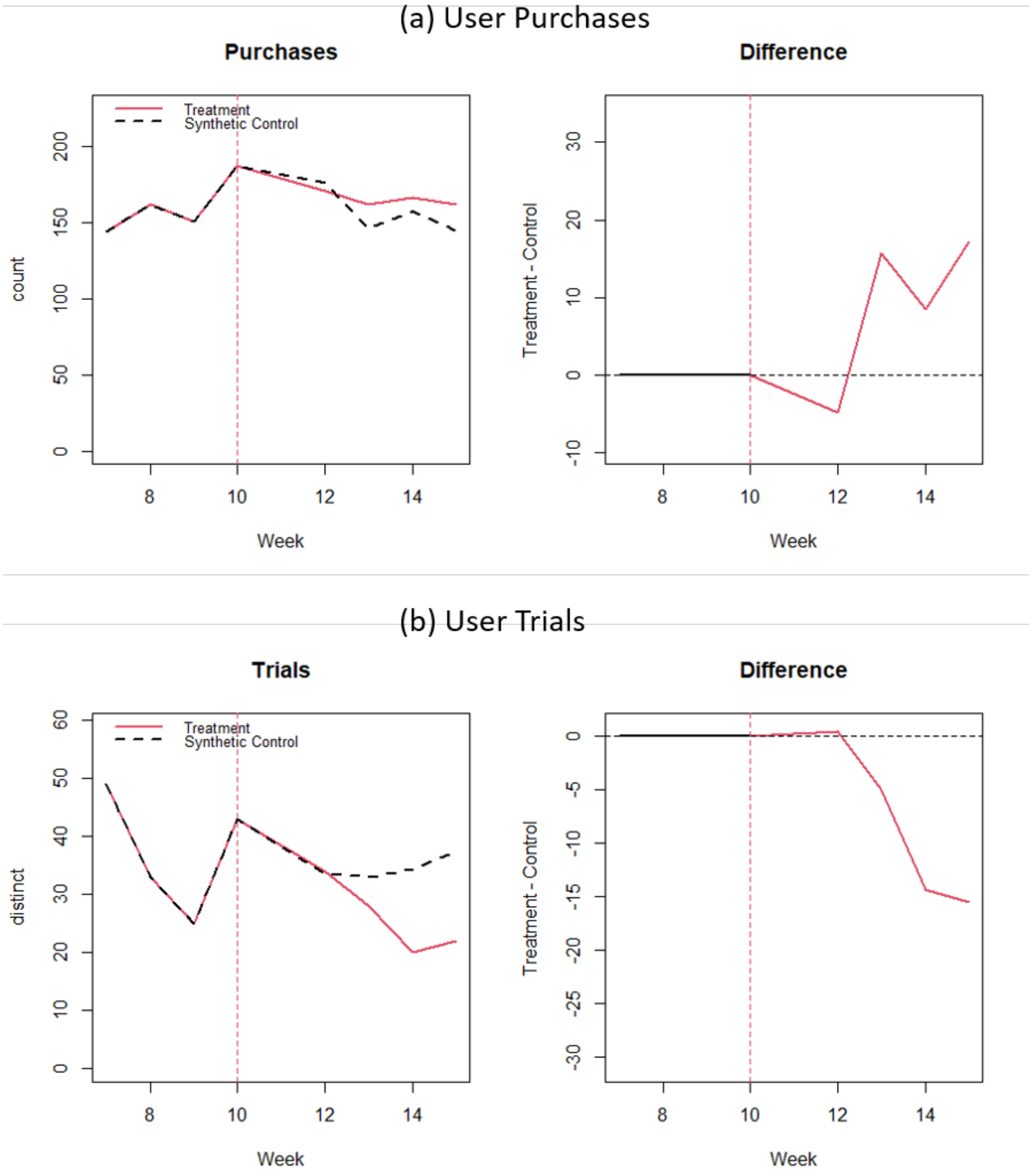


Figure 9 **(a) Synthetic control for user purchases. The treatment and synthetic control units match well in the pre-experiment period, and purchases increase for the treatment units after the experiment.**

(b) Synthetic control for user trials. The treatment and synthetic control units match well in the pre-experiment period, and trials decrease for the treatment units after the experiment.

7.4.3. Mechanism Check The experiment exogenously induced rating inflation in the treated restaurants. Consumers in that neighborhood were exposed to inflated ratings and responded by marginally increasing purchases and significantly decreasing trial. However, it is possible that restaurants also responded to the experiment and made changes to their offerings. If this were the case, we cannot be certain whether our results reflect changes in consumer behavior or changes made by restaurants in response to rating inflation. In this section, we describe the reasons why the observed results are likely driven by consumer responses rather than restaurant responses to rating inflation.

First, to rule out potential responses by the restaurant, we conduct additional analysis on the restaurants by checking whether the average amount spent per transaction from a restaurant changes due to the experiment. If the restaurant were to raise their prices or make changes to their menu, we could expect to see a change in the average amount spent per transaction at the treated restaurants. However, we find that there is no statistically significant difference in the average amount spent per transaction at treated restaurants as shown in Table 9.

Table 9 Average amount spent per transaction does not change

DV = *avg_bill*	(1)	(2)	(3)
D_{st}	-6.047	-4.361	-4.322
	(9.229)	(4.700)	(4.695)
Week Fixed Effects	Yes	-	Yes
Restaurant Fixed Effects	-	Yes	Yes
Observations	26927	26924	26924

Cluster-robust standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Further, we note that our post-experiment period is about a month, which is a relatively short duration. While the restaurant can potentially change their pricing and menu options in response to rating inflation, these changes are likely to take a longer time period to manifest. We also note that all the restaurants on the platform have a physical presence as a brick-and-mortar restaurant, with only a fraction of their total sales being generated through the food delivery app. Therefore, we believe the change in ratings on a food delivery app is unlikely to cause them to change their pricing or menu options, at least in the short term. Finally, we note that many restaurants are listed on other food delivery apps as well, which further diminishes the probability of them making significant changes based on the ratings system of one food delivery app. The empirical result as well as these arguments boosts our confidence that the observed effects are unlikely to be driven by restaurant responses of changes in price or menu.

8. Discussion and Conclusion

We study the consequences of rating inflation in a quasi-experiment setting. The results lead to several important insights for managers, designers, and developers of digital platforms that use ratings to help users choose between numerous sellers.

First, rating inflation should be viewed not just as an increase in average ratings, but also as a decrease in the informativeness of ratings due to lowering of the variance across ratings. Both these factors influence users in different ways. High average ratings lead to greater user purchases, while lower rating variance reduces how much users trial new restaurants. Combined, these two effects lead to greater sales concentration as sales shift towards more popular restaurants.

Second, rating inflation can potentially hurt platform growth in two ways:

a) Rating inflation may hurt the platform by reducing the informativeness of the rating system. Facilitating discovery and trial of new restaurants is an important component of the value proposition of digital platforms. Ratings reduce the search cost to find quality restaurants by allowing the consumer to leverage and learn from other's experiences. When rating inflation reduces the extent of trials by users, it erodes an important source of the value it provides to users: facilitating discovery and trial.

b) Increased sales concentration due to rating inflation can hurt platform growth by reducing the market power of the platform relative to popular sellers on the platform. Most platforms negotiate the terms of a seller's participation on the platform based on their relative market power. A rating system that increases sales concentration would harm the market power of the platform relative to the most popular sellers. In the context of this study, the participating platform negotiates the commission it receives from each restaurant per order based on the restaurant's relative market power. The most popular restaurants pay little or no commission to the platform while the less popular restaurants pay a larger commission. With these considerations, platforms have an incentive to reduce excess sales concentration on the platform.

Overall, our findings demonstrate that managers and designers need to account for and strategize to minimize rating inflation on their platform to ensure its health and growth. Our discussions with the managers of the platform partners for this study revealed that managers are cognizant of the possibility of the strategic implications of rating inflation. As one consequence of this study, the platform decided to remove the rating deflation they had imposed to all users in the focal city, and eventually nationally as they prioritized increasing purchases at their stage of growth.

Our work contributes to the growing stream of the Information Systems literature that has used randomized, natural, or quasi-experiments to uncover the mechanisms by which online ratings and review systems influence consumers (Burtch et al. 2018, Huang et al. 2019a,b, Shukla et al. 2021).

While this study identifies and expands our understanding of the impacts of rating inflation, it has a few limitations. First, since our observation period extends one month after the rating change, we are unable to investigate the long-term effects of rating inflation. The effects may attenuate over time as users recalibrate their expectations of what a high rating and a low variance represent, as compared to their outside option (Ho et al. 2017). Second, our partner platform does not collect textual reviews to maintain ease-of-use of their platform. As such, we are unable to measure if textual reviews alleviate the problem of rating inflation. Third, the role of culture in how users respond to rating inflation cannot be determined through this study as we have data from a single city. These areas can be promising directions for future research.

We also note the limits of generalizability of our findings given the specific institutional context in which the experiment was performed. First, while rating inflation in practice usually manifests endogenously and gradually over time, rating inflation in our study happens exogenously at once due to the platform's action. Therefore, the results from the current experiment may not incorporate factors in practice that endogenously drive rating inflation. We also note that here was a temporary platform-wide disruption during the roll-out of the experiment which we have attempted to account for in our analysis by dropping data from the week of the experiment roll-out. The current experiment setting is still valuable because it provides a unique and feasible opportunity to understand the direct influence of an exogenous change in rating mean and variance, without being confounded by endogenous causes of rating inflation in practice. The current experiment setting, although challenging and costly to conduct for the platform, exogenously induces rating inflation so that we could get causal estimates of the effect. It serves as the first step towards understanding the effect of rating inflation in practice.

Finally, the effect of rating inflation may be different under two scenarios: 1) the platform does not curate the sellers (e.g., Yelp) and rating inflation happens by changing ratings only, without changing the types of sellers on the platform; 2) the platform curates the sellers and rating inflation is driven by changing the types of sellers on the platform. In our context, although the platform has the ability to strategically curate the sellers in the long run, it did not change the selection of the sellers during our sample period. Therefore, our results represent the scenario in which rating inflation is only driven by changes in the ratings, given the same set of sellers. Our results may not be generalizable to the case in which rating inflation is caused by platforms' strategically selection of the sellers; this would be an interesting avenue for future research.

References

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

Adamopoulos P, Ghose A, Todri V (2018) The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research* 29(3):612–640.

Ai C, Norton EC (2003) Interaction terms in logit and probit models. *Economics letters* 80(1):123–129.

Anderson M, Magruder J (2012) Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122(563):957–989.

Aral S (2014) The problem with online ratings. *MIT Sloan Management Review* 55(2).

Azoulay P, Zivin JSG, Wang J (2010) Superstar extinction. *The Quarterly Journal of Economics* 125(2):549–589.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119(1):249–275.

Bolton G, Greiner B, Ockenfels A (2013) Engineering trust: reciprocity in the production of reputation information. *Management Science* 59(2):265–285.

Burtch G, Hong Y, Bapna R, Griskevicius V (2018) Stimulating online reviews by combining financial incentives and social norms. *Management Science* 64(5):2065–82.

Cabral L, Hortacsu A (2010) The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics* 58(1):54–78.

Cai H, Chen Y, Fang H (2009) Observational learning: Evidence from a randomized natural field experiment. *American Economic Review* 99(3):864–82.

Cameron AC, Trivedi PK (2013) *Regression analysis of count data. Vol. 53* (Cambridge university press,).

Cenfetelli RT, Schwarz A (2011) Identifying and testing the inhibitors of technology usage intentions. *Information systems research* 22(4):808–823.

Center PR (2016) Online shopping and e-commerce.

Chen PY, Hong Y, Liu Y (2018) The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science* 64(10):4629–4647.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43(3):345–354.

Ciani E, Fisher P (2018) Dif-in-dif estimators of multiplicative treatment effects. *Journal of Econometric Methods* 8(1).

Correia S (2015) Singletons, cluster-robust standard errors and fixed effects: A bad mix. Technical Note, Duke University.

Correia S, Guimarães P, Zylkin T (2019) Verifying the existence of maximum likelihood estimates for generalized linear models. arXiv preprint arXiv:1903.01633.

Correia S, Guimarães P, Zylkin T (2020) Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal* 20(1):95–115.

Dai W, Jin G, Lee J, Luca M (2018) Aggregation of consumer ratings: An application to yelp.com. *Quantitative Marketing and Economics* 16(3):289–339.

Dellarocas C, Wood CA (2008) The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science* 54(3):460–476.

Dimoka A, Hong Y, Pavlou PA (2012) On product uncertainty in online markets: Theory and evidence. *MIS quarterly* 395–426.

Duan W, Gu B, Whinston AB (2008a) Do online reviews matter? - an empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Duan W, Gu B, Whinston AB (2008b) The dynamics of online word-of-mouth and product sales - an empirical investigation of the movie industry. *Journal of Retailing* 84(2):233–242.

Fang L (2022) The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science* .

Feldman JM, Lynch JG (1988) Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of applied Psychology* 73(3):421.

Fernández-Val I, Martin W (2016) Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics* 192(1):291–312.

Filippas A, Horton JJ, Golden J (2018) Reputation inflation. *Proceedings of the 2018 ACM Conference on Economics and Computation.* (ACM).

Fradkin A, Grewal E, Holtz D (2018) The determinants of online review informativeness: Evidence from field experiments on airbnb. Working Paper.

Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. *Marketing Science* 31(3):448–473.

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

Hong YK, Pavlou PA (2014) Product fit uncertainty in online markets: Nature, effects, and antecedents. *Information Systems Research* 25(2).

Huang N, Burtch G, Gu B, Hong Y, Liang C, Wang K, Fu D, Yang B (2019b) Motivating user-generated content with performance feedback: Evidence from randomized field experiments. *Management Science* 65(1):327–345.

Huang N, Burtch G, Hong Y, Polman E (2016) Effects of multiple psychological distances on construal and consumer evaluation: A field study of online reviews. *Journal of Consumer Psychology* 26(4):474–482.

Huang N, Sun T, Chen Py, Golden JM (2019a) Word-of-mouth system implementation and customer conversion: A randomized field experiment. *Information Systems Research* 30(3):805–818.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Khurana S, Qiu L, Kumar S (2019) When a doctor knows, it shows: An empirical analysis of doctors' responses in a qa forum of an online healthcare portal. *Information Systems Research* 30(3):872–891.

Kokkodis M (2019) Reputation deflation through dynamic expertise assessment in online labor markets. *In The World Wide Web Conference*, 896–905.

Lewis G, Zervas G (2016) The welfare impact of consumer reviews: A case study of the hotel industry. Working Paper.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Information Systems Research* 19(4):456–474.

Lu X, Ba S, Huang L, Feng Y (2013) Promotional marketing or word-of-mouth? evidence from online restaurant reviews. *Information Systems Research* 24(3):596–612.

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104(8).

Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

Nosko C, Tadelis S (2015) The limits of reputation in platform markets: An empirical analysis and field experiment. *National Bureau of Economic Research* (w20830).

Proserpio D, Xu W, Zervas G (2018) You get what you give: theory and evidence of reciprocity in the sharing economy. *Quantitative Marketing and Economics* 16(4):371–407.

Puhani PA (2012) The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters* 115(1):85–87.

Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

Shang S, Nesson E, Fan M (2018) Interaction terms in poisson and log linear regression models. *Bulletin of Economic Research* 70(1):89–96.

Shukla AD, Gao G, Agarwal R (2021) How digital word-of-mouth affects consumer decision making: Evidence from doctor appointment booking. *Management Science* 67(3):1329–1992.

Silva JS, Tenreyro S (2006) The log of gravity. *The Review of Economics and Statistics* 88(4):641–658.

Silva JS, Tenreyro S (2011) Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator. *Economics Letters* 112(2):220–222.

Song T, Huang J, Tan Y, Yu Y (2019) Using user- and marketergenerated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research* 30(1):191–203.

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

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

Wang CA, Zhang XM, Hann IH (2018) Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. *Information Systems Research* 29(3):641–655.

Wang K, Goldfarb A (2017) Can offline stores drive online sales? *Journal of Marketing Research* 54(5):706–719.

Wooldridge JM (1997) Quasi-likelihood methods for count data. *Handbook of Applied Econometrics* 2:352–406.

Wooldridge JM (2010) Econometric analysis of cross section and panel data. *MIT press* .

Wu C, Che H, Chan TY, Lu X (2015) The economic value of online reviews. *Marketing Science* 34(5):739–754.

Yi C, Jiang Z, Benbasat I (2017) Designing for diagnosticity and serendipity: An investigation of social product-search mechanisms. *Information Systems Research* 28(2):413–429.

Zervas G, Proserpio D, Byers J (2020) A first look at online reputation on airbnb, where every stay is above average. SSRN.

Zhao Y, Yang S, Narayan V, Zhao Y (2013) Modeling consumer learning from online product reviews. *Marketing Science* 32(1):153–169.

Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing* 74(2):133–148.

Appendix A: Review of Rating Systems

The design of ratings by platforms is a significant factor in influencing how users perceive and interpret these ratings (Acemoglu et al. 2017). Most platforms use some form of statistical aggregation of user feedback into a quality metric (Dai et al. 2018, Nosko and Tadelis 2015); and the choice of how ratings are aggregated and displayed determines how informative ratings are for consumers. We provide a brief overview of different types of rating systems in common use to contrast their differences.

Consider the Michelin star rating of restaurants.[2] A select few restaurants are awarded 1, 2 or 3 stars every year based on reviews by professional food critics. These rating signals are highly prestigious and there is anecdotal evidence that a difference of one star in a restaurant's rating has a significant impact on their business.[34] Yet, consumers cannot use the Michelin star rating to distinguish between the quality of restaurants with identical star-ratings. In addition, restaurants that are not awarded a star are also not differentiated.

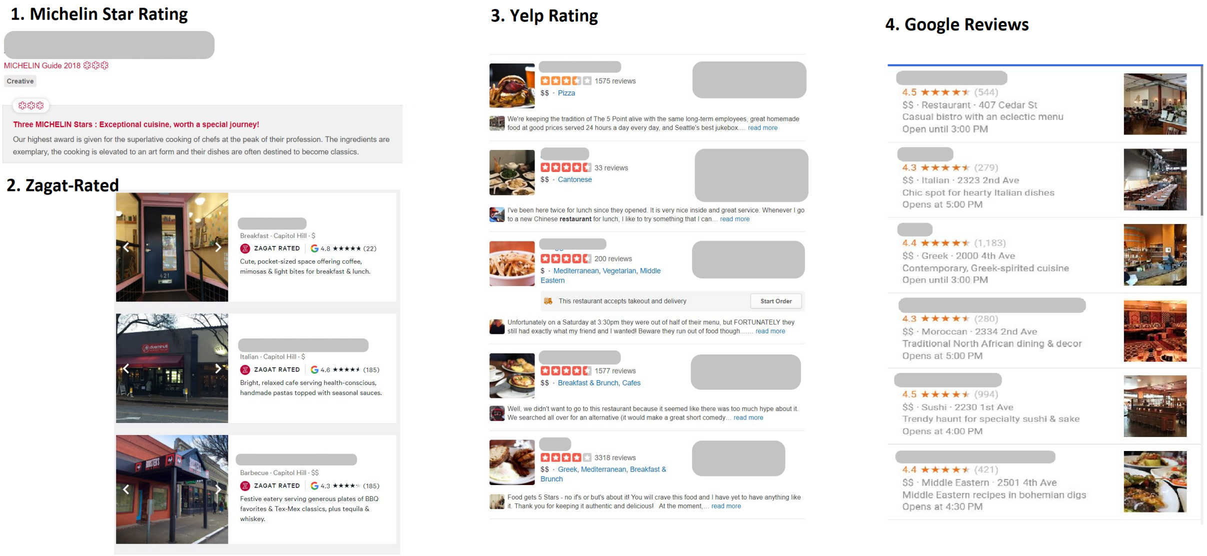
Yelp, a popular review website, displays aggregated ratings which are rounded-off to a half-star out of five stars. Thus there are ten distinct tiers for restaurants. In addition, Yelp also displays the number of ratings each restaurant has received. The stars are color-coded in different shades of orange to red to help users visually identify top rated restaurants. On the restaurant page, more information about ratings, including the distribution of ratings given by users, as well as their textual reviews and photographs are provided.

Google also collates feedback from users into a consolidated rating. The scale used is again 5-stars, but ratings are rounded-off to the nearest one decimal place. The lowest rating that can be given is a 1-star. When aggregated, the rating can vary from 1.0 to 5.0 in increments of 0.1 star and thus there are 40 possible tiers that a restaurant can belong to. Given the greater number of possible ratings a restaurant can have, Google's rating enables users to differentiate more restaurants than Yelp's rating.

[2] Michelin Guides https://en.wikipedia.org/wiki/Michelin_Guide, Retrieved on June 1, 2018

[3] 'Michelin blessing could mean 25% bump for one-star restaurants' http://www.chicagobusiness.com/article/20101129/NEWS07/101129946/michelin-blessing-could-mean-25-bump-for-one-star-restaurants, Retrieved June 1, 2018

[4] 'The Impact of Michelin Stars on Business' https://www.thestaffcanteen.com/Editorials-and-Advertorials/ impact-michelin-stars-business, Retrieved June 1, 2018



1. Michelin Star Rating

[redacted]

MICHELIN Guide 2018 ✿✿✿

Creative

✿✿✿

Three MICHELIN Stars : Exceptional cuisine, worth a special journey!

Our highest award is given for the superlative cooking of chefs at the peak of their profession. The ingredients are exemplary, the cooking is elevated to an art form and their dishes are often destined to become classics.

2. Zagat-Rated

[redacted]

Breakfast · Capitol Hill · $

ZAGAT RATED | 4.8 ★★★★★ (22)

Cute, pocket-sized space offering coffee, mimosas & light bites for breakfast & lunch.

[redacted]

Italian · Capitol Hill · $

ZAGAT RATED | 4.6 ★★★★★ (185)

Bright, relaxed cafe serving health-conscious, handmade pastas topped with seasonal sauces.

[redacted]

Barbecue · Capitol Hill · $$

ZAGAT RATED | 4.3 ★★★★☆ (185)

Festive eatery serving generous plates of BBQ favorites & Tex-Mex classics, plus tequila & whiskey.

3. Yelp Rating

[redacted]

★★★★☆ 1575 reviews

$$ · Pizza

[redacted]

We're keeping the tradition of The 5 Point alive with the same long-term employees, great homemade food at good prices served 24 hours a day every day, and Seattle's best jukebox.... read more

[redacted]

★★★★★ 33 reviews

$$ · Cantonese

[redacted]

I've been here twice for lunch since they opened. It is very nice inside and great service. Whenever I go to a new Chinese **restaurant** for lunch, I like to try something that I can... read more

[redacted]

★★★★★ 200 reviews

$ · Mediterranean, Vegetarian, Middle Eastern

[redacted]

This restaurant accepts takeout and delivery | Start Order

Unfortunately on a Saturday at 3:30pm they were out of half of their menu, but FORTUNATELY they still had exactly what my friend and I wanted! Beware they run out of food though....... read more

[redacted]

★★★★★ 1577 reviews

$$ · Breakfast & Brunch, Cafes

[redacted]

Well, we didn't want to go to this restaurant because it seemed like there was too much hype about it. We searched all over for an alternative (it would make a great short comedy... read more

[redacted]

★★★★☆ 3318 reviews

$$ · Greek, Mediterranean, Breakfast & Brunch

[redacted]

Food gets 5 Stars - no if's or but's about it! You will crave this food and I have yet to have anything like it. Thank you for keeping it authentic and delicious!   At the moment,... read more

4. Google Reviews

[redacted]

4.5 ★★★★★ (544)

$$ · Restaurant · 407 Cedar St

Casual bistro with an eclectic menu

Open until 3:00 PM

[redacted]

4.3 ★★★★★ (279)

$$ · Italian · 2323 2nd Ave

Chic spot for hearty Italian dishes

Opens at 5:00 PM

[redacted]

4.4 ★★★★★ (1,183)

$$ · Greek · 2000 4th Ave

Contemporary, Greek-spirited cuisine

Open until 3:00 PM

[redacted]

4.3 ★★★★★ (280)

$$ · Moroccan · 2334 2nd Ave

Traditional North African dining & decor

Opens at 5:00 PM

[redacted]

4.5 ★★★★★ (994)

$$ · Sushi · 2230 1st Ave

Trendy haunt for specialty sushi & sake

Opens at 4:00 PM

[redacted]

4.4 ★★★★★ (421)

$$ · Middle Eastern · 2501 4th Ave

Middle Eastern recipes in bohemian digs

Opens at 4:30 PM

Figure 10 **Different Restaurant Rating systems. 1) Michelin Star Rating, 2) Yelp Ratings, 3) Google Reviews**

Appendix B: New Users

In the main analysis of the paper, we excluded users who joined the platform within one month prior to the experiment or after the experiment. In this section, we present results from the sub-sample of new users who join the platform during the observation period. We compare the purchases for users who join after the experiment with that of those who join before. Specifically, we estimate the following equation:

$$n\_purchase_{it} \sim \text{Poisson}\Big[\exp[\beta_1 join\_after_i + \beta_2 treated_i + \beta_3 join\_after_i \,\text{x}\, treated_i + \tau_t)\Big] \tag{7}$$

where $n\_purchase_{it}$ is the count of purchases for user i in week t, $join\_after_i$ is an indicator variable equal to 1 if the user has joined the platform after the experiment, and 0 otherwise. $treated_i$ is the indicator variable equal to 1 if the user belongs to the treated group, and 0 otherwise. tau_t is the weekly fixed effects.

Table 10 Purchases for new users joining before or after the experiment

	(1) *n_purchase*
join_after	1.168***
	(0.0135)
treated	-0.115***
	(0.0187)
join_after x *treated*	0.173***
	(0.0354)
constant	-0.905***
	(0.00488)
Week Fixed Effects	Yes
Observations	544463

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The results are shown in Table 10. We find that:

- users who joined the platform after the experiment in the treated neighborhood purchased more than those who joined before the experiment in the treated neighborhood, suggesting that "joining after the experiment" is associated with more purchase, and
- users who joined the platform after the experiment in the treated neighborhood purchased more than those who joined after the experiment in the control neighborhood, suggesting that "joining in the treated neighborhood" is associated with more purchase.

Both results are consistent with the finding of our main model that the experiment (i.e., rating inflation) can cause a higher *n_purchase*.

Appendix C: Additional Robustness Checks

C1: User purchase frequency thresholds

In all of our analysis in the paper, we dropped users who had more than three purchases from a single restaurant on one day. These are likely to be group accounts shared by multiple users, and hence may behave differently than individual user accounts. In this section, we vary the user purchase frequency cutoff to include: (i) all users, (ii) users with less than 5 purchases from a single restaurant on one day. The results are reported in Table 11. We find these results are consistent with the main results.

Table 11 Robustness checks including all users (columns 1 & 3), and by dropping users who had more than 5 purchases from the same restaurant on a single day (columns 2 & 4).

	n_purchase		*n_trial*	
	All users	Freq < 5	All users	Freq < 5
	(1)	(2)	(3)	(4)
D_{st}	0.000424	0.0125	-0.076**	-0.074**
	(0.0202)	(0.0203)	(0.0299)	(0.0300)
Week Fixed Effects	Yes	Yes	Yes	Yes
User Fixed Effects	Yes	Yes	Yes	Yes
Observations	1429209	1418670	563132	556505

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

C2: Pre-Treatment Placebo Tests

In this section, we perform pre-treatment placebo tests as further robustness checks of our main results. We limit our sample to the pre-experiment period and assign the duration between 8 and 10 weeks as the treatment period as a placebo for users, and between 6 and 10 weeks as the placebo for restaurants. Since there was no actual treatment during this period, the results should be insignificant, which is the case as shown in Table 12. This provides further evidence that our main results are driven by the experiment and not due to temporal trends that differ between the treated and control neighborhoods.

Table 12 Pre-Treatment Placebo Tests

	Users		Restaurants	
	n_purchase	*n_trial*	*sales*	*sales_conc*
	(1)	(2)	(3)	(4)
D_{st}	-0.0706	-0.0529	0.674	0.0041
	(0.0536)	(0.0347)	(1.310)	(0.057)
Week Fixed Effects	Yes	Yes	Yes	Yes
User Fixed Effects	Yes	Yes	-	-
Restaurant Fixed Effects	-	-	Yes	Yes
Observations	676840	276503	20214	18035

Cluster-robust standard errors in parentheses

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

Appendix D: Neighborhood Total Plots

We show the neighborhood-level total user purchases and trials by group in Figure 11.

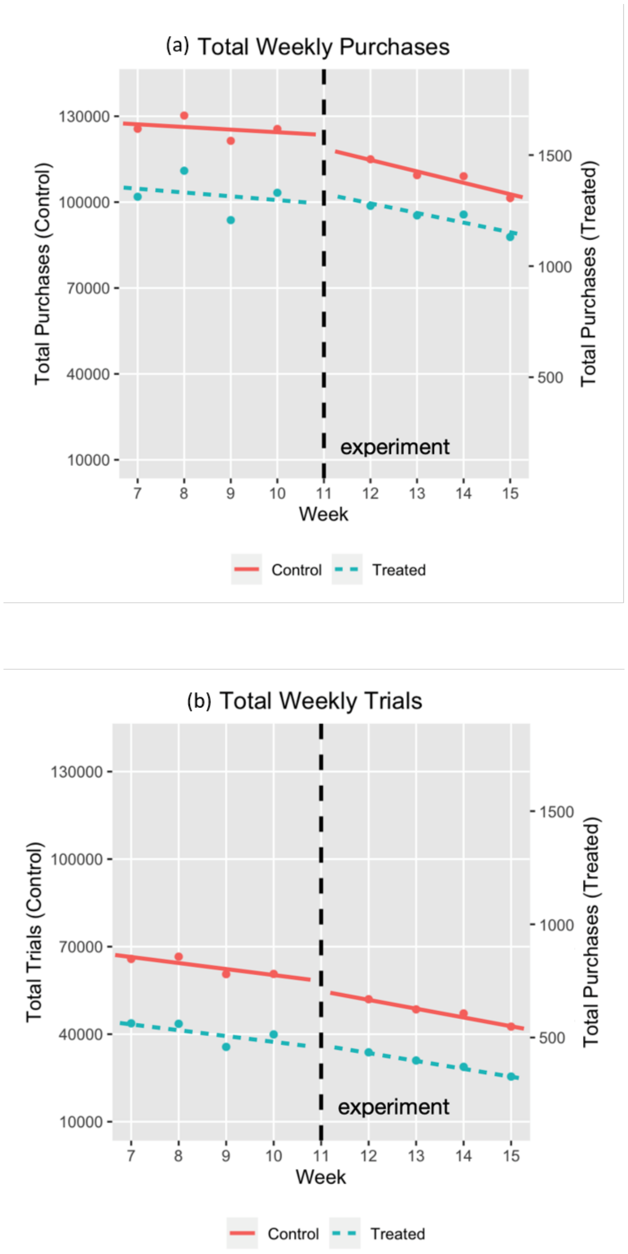


Figure 11 **(a) Total user purchases by group. Total weekly purchases by treated and control users are parallel before the experiment.**

(b) Total user trials by group. Total weekly trials by treated and control users are parallel before the experiment.