Check for updates

Article

Picky Eaters Make for Better Raters

Cornell Hospitality Quarterly
2025, Vol. 66(1) 85–94
© The Author(s) 2024

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/19389655241226557
journals.sagepub.com/home/cqx

Sasha Stoikov[1], Stefano Borzillo[2], and Steffen Raub[2]

Abstract

It has been established in the literature that the number of ratings and the scores restaurants obtain on online rating systems (ORS) significantly impact their revenue. However, when a restaurant has a limited number of ratings, it may be challenging to predict its future performance. It may well be that ratings reveal more about the user who gave the rating than about the quality of the restaurant. This motivates us to segment users into "inflating raters," who tend to give unusually high ratings, and "deflating raters," who tend to give unusually low ratings, and compare the rankings generated by these two populations. Using a public dataset provided by Yelp, we find that deflating raters are better at predicting restaurants that will achieve a top rating (4.5 and above) in the future. As such, these deflating raters may have an important role in restaurant discovery.

Keywords

guest ratings; online rating systems; restaurant industry; ratings inflation

Introduction

Online rating systems (ORSs) have transformed how consumers decide what items to purchase, movies to watch, hotels to stay in, wine to drink, and restaurants to eat at. In theory, by democratizing the act of criticism, ORSs have challenged gatekeepers of the past and opened a path of discovery to a myriad of products and services. In practice, however, the last 20 years have seen the rise of superstars that have dominated these ORSs and made it harder for new entrants to gain traction. Restaurants are no exception to this phenomenon.

Can restaurant ratings be trusted? Answering this seemingly simple question requires considering the stakeholders involved in the rating process, that is, restaurants, users, and ORSs. Each has incentives to inflate their rating. First, restaurants interested in increasing their reputation may be tempted to inflate their scores through fraud or questionable methods (Luca & Zervas, 2015; E. Y. Wang et al., 2022; Zhang et al., 2022; Zhu et al., 2019). Second, users fulfill their tendency to rate more things they like than things they dislike (Dey & Kumar, 2023; Park & Nicolau, 2015; Stamolampros & Korfiatis, 2018). Finally, ORSs may benefit from the fact that high ratings lead to more consumption (Aziz et al., 2023). The actions of all of these stakeholders lead to inflated ratings which may or may not be deserved. Inflated rates thus make it hard to distinguish between truly exceptional restaurants and those that benefit from inflated ratings.

To distinguish the truly exceptional restaurants, there is a need for more critical opinions. Like Michelin critics, users who are hard to please—which we refer to as "deflating raters"—may be better suited to uncover restaurant quality than the average rater. It is possible that within an ORS, a subset of "deflating" raters may deflate the restaurants whose ratings are less deserved and boost the reputations of lesser-known gems. We are interested in identifying these deflating raters and demonstrating their value in predicting future restaurant scores.

In the domain of restaurant ratings, which are fundamentally linked to the taste of the users, the role of user heterogeneity is likely to be important. Not only do users disagree on what makes a good or a bad restaurant experience, but they differ on the statistics of their ratings. In this article, the statistic we focus on is the average rating per user, which indicates how "picky" that user is. This article will define "inflating raters" as the users who give unusually high ratings and "deflating raters" as those who give unusually low ratings. Intuitively, a 4-star rating does not mean the same thing to a deflating rater with an average rating of 3 stars as

[1]Cornell Financial Engineering Manhattan, New York, NY, USA
[2]EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland

Corresponding Author:
Sasha Stoikov, Cornell Financial Engineering Manhattan, New York, NY, USA,
Email: sfs33@cornell.edu

to an inflating rater with an average rating of 5 stars. Therefore, every rating is driven by both the restaurant's quality and the user's average rating. In this context, our first research objective is thus to understand whether restaurant ratings say more about the user who gave the rating or about the quality of the restaurant.

After highlighting the importance of the users and segmenting them into inflating and deflating raters, we were interested in making a value judgment on these two populations. Thus, our second research objective was to determine whether deflating raters are better than inflating raters. To answer this question, we define what makes a good user in terms of their ability to predict future scores. Since restaurants with a high number of ratings and high scores have been shown to perform better in terms of revenue, this is a good proxy for the future performance of restaurants.

Literature Review

Economic Impact of Ratings

Why are online ratings important to the economy? An important strand of the literature on restaurant and hotel-related ORSs focuses on the impact of ratings on the economic performance of the businesses being rated. Combining ratings data from ORSs and business performance data (from public records or directly from ORSs), extant research has established a strong correlation between the two. Luca (2016), Aziz et al. (2023), and Sayfuddin and Chen (2021), for instance, found that higher-rated businesses are associated with higher revenues. Studies have also shown that other performance-related metrics such as net sales, guest count, and average check (W. G. Kim et al., 2016), reservation availability (Anderson & Magruder, 2012), profitability (Y. Wang et al., 2021), transaction size (Torres et al., 2015), and restaurant closure (Yoo & Suh, 2022) are affected by ratings. What emerges from all these studies is that both the quantity of ratings and average score for a business positively impact the business's bottom line.

Review Usefulness

What makes a single review useful? An important feature of most ORSs is that they give users the ability to rate the usefulness of reviews. After reading a review, a user may respond to the question, "Was this review helpful?" An extensive empirical literature using Yelp and TripAdvisor reviews has emerged around identifying features that explain what constitutes a useful or helpful review.

The first feature to attract attention was the numerical value of the rating itself. H. Lee et al. (2011) found that negative reviews were typically deemed more useful than positive reviews. This has been confirmed and refined by subsequent multiple studies (Filieri et al., 2019; Kwok & Xie, 2016; Luo et al., 2020; Mudambi & Schuff, 2010; Salehi-Esfahani & Ozturk, 2018; Yang et al., 2017; Zhu et al., 2019). In a somewhat similar vein, Park and Nicolau (2015) have shown that extreme ratings, that is, 1 or 5 stars, are considered more useful than moderate ratings.

A second feature authors focused on is the content of the review to uncover what constitutes a helpful review. Using text mining approaches to uncover the emotional aspects of the review text, H. Li et al. (2020) and X. Wang et al. (2019), for instance, found that reviews conveying negative emotions are more useful than those containing positive emotions. On their side, Z. Liu and Park (2015) introduced metrics based on qualitative aspects of the review text, such as the readability of the text. They found such metrics to be most influential on the review's usefulness. Just as interestingly, Ma et al. (2018) studied the inclusion of photos in the review and found them helpful when combined with the review text.

Third, extant studies have also widely reported that features related to the reviewer are important in assessing reviews' usefulness. These studies have focused on the reviewer's specific features, such as elite status (Hlee et al., 2019; Kwok & Xie, 2016), years of experience on the ORS (Kwok & Xie, 2016), number of cities visited (Kwok & Xie, 2016; Liang et al., 2019), and number of followers (Hlee et al., 2019; M. Lee et al., 2021). In general, the better the reputation of a reviewer, the more useful other users deem the review.

Two recurrent themes (related to the reviewer) emerge from what has been described earlier: the importance of negative opinions and the reputation of the reviewer. This suggests that the most useful reviews on Yelp and TripAdvisor come from critical, experienced reviewers who are sparing with their positive reviews. These users, much like Michelin critics, may be most qualified to identify true gems in the restaurant landscape.

Bias in the Rating Process

There has been a significant interest in the biases that can affect the reliability of ratings given by online consumers (e.g., G. Gao et al., 2015; Han & Anderson, 2020; Hu et al., 2017; Jiang & Guo, 2015; X. Li & Hitt, 2008; F. X. Wang & Anderson, 2023). Among these biases, the self-selection bias is particularly prominent, as participating in ORS is a fundamentally voluntary activity. Self-selection bias occurs when a specific category of consumers is more inclined to write reviews. This phenomenon may lead to a distortion in the representation of average opinions, with published reviews failing to capture the full range of experiences of all users. Hu et al. (2017) have broken down the self-selection bias into two distinct biases: the underreporting bias and the acquisition bias.

These two biases play a key role in explaining the J-shaped distribution of online product reviews. Underreporting bias contributes to the bimodality of the distribution while acquisition bias is responsible for the asymmetry and tendency to inflate ratings.

Underreporting bias occurs when consumers with moderate or average experiences are less likely to write reviews, leading to a lack of representativeness of "ordinary" experiences in overall reviews. The underreporting bias in online reviews, which reflects the underrepresentation of moderate experiences, is closely related to the propensity to rate bias identified by Han and Anderson (2020). The latter points out that extreme, particularly negative reviews are more frequently published, especially by users less accustomed to the review process. Together, these biases explain the tendency of online reviews to lean toward more extreme ratings, accentuating the predominance of negative reviews, especially for consumers unfamiliar with review platforms. This motivates our focus on raters with some experience, that is, more than 5 ratings, on the Yelp platform.

Acquisition bias occurs when consumers with positive expectations or exceptional experiences are more inclined to acquire a product or service and, consequently, to leave reviews. This can often lead to inflated ratings, as only consumers with a positive expected net utility tend to write a review (Hu et al., 2017). However, this tendency toward inflated ratings can be ambiguous: These reviews may reflect an authentic and faithful assessment of the consumer experience or overconfidence in their ability to choose good products or services. It is important to recognize that inflated ratings are prevalent in the context of online reviews, and their potential detrimental impact has been widely documented.

Inflated Ratings: Why They Are Detrimental and How They Can Be Mitigated

It has been widely observed that ORSs have implausibly high ratings, and researchers have expressed concern that it may be difficult to distinguish between individuals or organizations that are truly exceptional and those that benefit from inflated ratings (Filippas et al., 2022; Garg & Johari, 2019b; Horton & Golden, 2015). Several empirical studies have established the prevalence of inflated ratings. On eBay, more than 90% of sellers studied between 2011 and 2014 had a rating of at least 98% positive (Nosko & Tadelis, 2015). On the online freelancing platform oDesk, average ratings rose by one star over 7 years (Filippas et al., 2022; Horton & Golden, 2015). On taxi platforms like Uber and Lyft, inflation is particularly extreme, as anything less than 5 stars puts a driver at risk of deactivation (Athey et al., 2019). On Airbnb, almost 95% of hosts have an average rating of 4.5–5 out of 5 stars (Zervas et al., 2017). On Amazon, ratings tend to be bimodal, with the biggest peak at 5 stars and a small peak at 1 star (Hu et al., 2017).

Recent research has looked at strategies for mitigating the occurrence of inflated ratings. A natural experiment on Bookings.com, where the rating system changed from a smiley rating to a 10-point rating, revealed a change in the distribution of ratings (J. M. Kim et al., 2023; Leoni & Boto-García, 2023), hinting at the role of the ratings interface on mitigating the occurrence of inflated ratings. There have been some proposed solutions to making ratings less inflated. First, it has been suggested to make feedback private (Horton & Golden, 2015). Second, a randomized control trial on an online labor market showed that positive-skewed verbal scales (putting words like "the best I have ever hired") significantly reduce the occurrence of inflated ratings (Garg & Johari, 2019a). Third, a randomized control trial on a music discovery platform showed that introducing timers on the higher end of the rating scale can encourage user pickiness (Shahout et al., 2023).

Our research focuses on a post-processing approach, where we select subsets of the user population who provide less-inflated ratings. In other words, we identify clusters of users who naturally submit lower average ratings. We then examine how well the ratings of these users predict restaurant quality.

Methodology

Dataset (Yelp)

We use data from the Yelp academic dataset for our analysis (see https://www.yelp.com/dataset). In particular, we use three provided tables: the review table, the business table, and the user table.

The review table has 7M reviews, out of which 4.7M reviews have the "restaurant" tag. These restaurant reviews have the following structure:

1. User_id
2. Business_id
3. Rating (1-5 star)
4. Other data, like review text, date, usefulness, and so on.

The business table has 52k businesses with the following structure:

1. Business_id
2. Rating count
3. Yelp Score (average rating rounded to the nearest half-integer)
4. Categories, which include the "restaurants" tag.
5. Other data, like location, hours, food type, and so on.

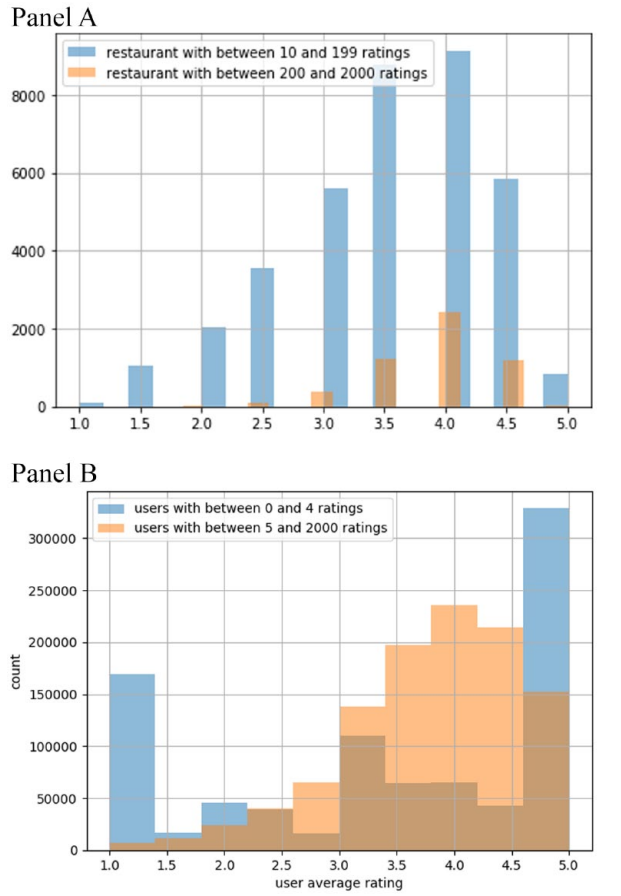

Figure 1.
Panel A: Yelp Score Distribution for Restaurants With Differing Quantity of Ratings. Panel B: Rating Distribution for Raters With Differing Quantity of Ratings.

The user table has 1.4M users with the following data structure:

1. User_id
2. Rating count
3. The average rating given by that user
4. Other data, like number of friends, user name, elite status, and so on.

To the best of our knowledge, the phenomenon of inflated ratings has not been investigated on Yelp though the average user rating in the dataset of 3.79 suggests that ratings tend to be higher than the objective quality of the restaurants, which we expect to be closer to 3, if the average restaurant is to be calibrated to the average score on the 1–5 scale. We constructed four figures for a better understanding of the Yelp dataset. In Figure 1, Panel A displays the Yelp score distribution for restaurants rated between 10 and 199, and restaurants rated between 200 and 2,000 times. Both distributions peak at a 4.0 Yelp score. For restaurants rated a modest amount of times, in the 10–199 range, there are a significant number of restaurants with a score of 3.0 or less. However, once a restaurant has been rated more than 200 times, the scores mostly fall within the 3.5–4.5 range. Notably, the Yelp dataset has approximately 5,000 restaurants with more than 200 ratings.

In Figure 1, Panel B displays the average rating distribution for users in the 0–4 ratings range and the 5–2,000 ratings range. We note that users with very few ratings produce ratings that are very polarized, with a majority of 5-star ratings and a large peak at 1-star ratings. After about five ratings, the distribution starts looking like the restaurant distribution: with a peak at 4 stars. This indicates that the underreporting bias is likely to be mitigated once a rater is familiar with the platform. However, the peak at 4 stars indicates that acquisition bias may still play an important role. We define "deflating" raters to be those in the bottom 25th percentile of users who rated five or more restaurants. Similarly, we define "inflating" raters as those in the top 75th percentile of users who rated five or more restaurants.

In Figure 2, Panel A shows the distribution of ratings from inflating raters, that is, users with an average rating greater than the 75th percentile threshold of 4.33. In Figure 2, Panel B shows the distribution of ratings from deflating raters, that is, users with an average rating lower than the 25th percentile threshold of 3.38. The distributions of inflating and deflating user's ratings are very different, with the deflating users producing an almost uniform distribution of ratings, while inflating users overwhelmingly dole out 5-star ratings. Inflating users have a high mean score and low variance. Deflating users have high variance and a mean centered around 3. As we will see in our findings in Discussion section, these statistics have significant downstream consequences on the ratings and rankings of restaurants.

Regression Analysis

Do ratings say more about the user who gave the rating or about the quality of the restaurant? We use linear regressions, similar to much of the literature on rating usefulness (described in the Review Usefulness subsection of our literature review) to explore this question. The main difference in our regression is that the dependent variable is the numerical value of the rating, rather than a rating usefulness metric. The independent variables are normalized average user rating and normalized restaurant Yelp score. The relative values of the coefficients indicate whether a single rating is more a product of the user's average behavior or the restaurant's average rating.

We expect the impact of the average user rating to get averaged out and disappear as a restaurant receives more

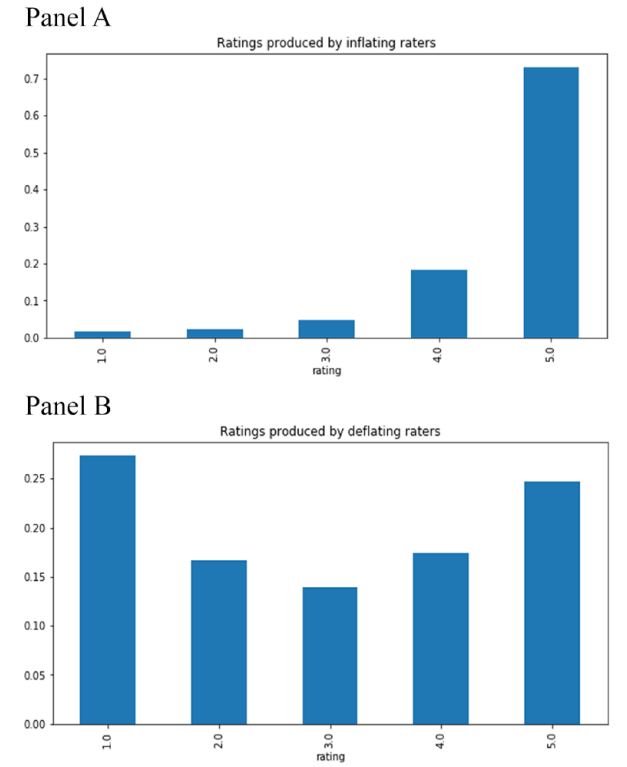

Figure 2.
Panel A: Ratings From Inflating Raters. Panel B: Ratings From Deflating Raters.

ratings. This motivates us to run a regression of average restaurant ratings (independent variable) against the average user rating for that restaurant (dependent variable), for restaurants rated in the 200–2,000 times range, to see if the relationship between average restaurant ratings and average user ratings persists.

This method differs from most of the literature on rating usefulness in that it involves averaging over multiple ratings on a per-restaurant basis. The two variables may be correlated, but the direction of causality is debatable. It may be that restaurants with higher ratings attract a higher proportion of inflating raters simply because these raters are better at finding quality restaurants. It is also plausible that restaurants with higher ratings are simply better and are motivating inflating raters to rate and find ways to motivate deflating raters to abstain. Indeed, one of the limitations of ratings datasets is that the true quality of a restaurant is not observable, making it challenging to assess the relative qualities of various types of raters.

In what follows, we will assume that for restaurants rated more than 200 times, the Yelp score is a proxy for quality, although the role of marketing in incentivizing inflating raters needs to be acknowledged.

Yelp Score Prediction

Our second research question concerns the relative value of raters. Are deflating raters better at predicting Yelp scores than inflating raters? Ideally, we want to know which group is more accurate in predicting the "true" quality ranking of the restaurants. Since this true restaurant quality is not observable, and arguably, does not even exist, we will need to use a proxy. In practice, we use the Yelp score (a simple average of all ratings for a restaurant, rounded to the nearest 0.5) for restaurants that have been rated more than 200 times and have been rated by at least 50 inflating raters and 50 deflating raters. The overwhelming majority of restaurants that fulfill these conditions have a rating of 3.5, 4.0, and 4.5, and we define our universe to be the 3,146 restaurants that fall in one of these three categories. We find 768 restaurants with a rating of 3.5 (22% of restaurants), 1,697 restaurants with a rating of 4 (54% of restaurants), and 681 restaurants with a rating of 4.5 (24% of restaurants). Thus, a random ranking of restaurants would have an accuracy of 22% in predicting restaurants with a 3.5 rating, 54% in predicting restaurants with a 4.0 rating, and 24% in predicting restaurants with a 4.5 rating.

The method we use generally falls under the family of bootstrapping methods. The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples. Samples are constructed by drawing observations from a large data sample with replacement. In our case, we are sampling from inflating and deflating raters to produce a ranking of our restaurant universe. This ranking is then used to classify restaurants into the low (3.5), middle (4.0), and high (4.5) categories and compared to the "true" ranking, given by the average of all ratings, that is, the Yelp score. Alternatively, we could exclude inflating and deflating raters from this average to compute the "true" ranking, but since the results were very similar, we focus on the Yelp score as a target to compute accuracy.

The entire process for bootstrapping, classifying restaurants, and the subsequent computation of accuracy scores can be summarized as follows:

1. For each restaurant in our universe and each user category (either "inflating" or "deflating" raters):
   a) Sample 20 users with replacement.
   b) Compute the average restaurant score.
   2. Sort the restaurant by average user score.
   3. Classify restaurants in the low (3.5), middle (4.0), and high (4.5) score categories, using percentile thresholds from the random baseline. In other words, the top 24% or restaurants are classified as 4.5, the next 54% are classified as 4.0, and the bottom 22% are classified as 3.5.
4. Compute the accuracy of this classification with respect to the Yelp score based on all ratings. In

Table 1.
Regression Analysis – Dependent Variable = Average User Rating of the Restaurant.

Independent Variables	Coefficient	*SE*	*t*-value	*p*-value
Intercept	3.7938	0.000	7598.20	.000
Average normalized user rating	0.6617	0.001	1281.539	.000
Average normalized restaurant rating	0.4214	0.001	816.131	.000
	R^2	0.391		
	Adj. R^2	0.391		
	F-value	1.520e+06		
	p-value	.000		
	df residuals	4,724,461		
	df model	2		

other words, the accuracy score represents the proportion of restaurants from the bootstrapping procedure that are correctly slotted into the score category to which they belong when all user ratings are considered.

By repeating the aforementioned procedure multiple times, we can compute an average accuracy as well as standard error bars. These error bars allow us to determine with confidence, whether deflating raters significantly outperform inflating raters.

Results

Regression Analysis

With the purpose of explaining user-restaurant ratings at the single rating level, we estimated the following equation:

$$y_{ij} = a + b_1 X_{1i}^{norm} + b_2 X_{2j}^{norm} + e_{ij}$$

where:

y_{ij} = rating by user i of restaurant j

$X_{1i} = \frac{1}{n_i}\sum_j y_{ij}$ = user i's average rating where n_i is the number of ratings by user i

$X_{2j} = \frac{1}{n_j}\sum_i y_{ij}$ = restaurant j's average n_j rating is the number of ratings for restaurant j

The normalized user and restaurant ratings are given by

$$X_{1i}^{norm} = \frac{X_{1i} - \mu_u}{\sigma_u} = \text{normalized user rating}$$

and

$$X_{2j}^{norm} = \frac{X_{2j} - \mu_r}{\sigma_r} = \text{normalized restaurant rating}$$

where μ_u and σ_u are the mean and standard deviations of all users' average ratings, and μ_r and σ_r are the mean and standard deviations of all restaurants' average ratings.

Results of the regression are displayed in Table 1.

The constant of this regression is 3.8, representing the average restaurant rating on Yelp. The normalized average user rating coefficient of 0.66 ($p < .001$) represents the impact of an increase or decrease of one standard deviation of the average user rating. This is significantly larger than the normalized restaurant rating average coefficient of 0.42 ($p < .001$), which represents the impact of an increase or decrease of one standard deviation of the restaurant score. The R squared of that regression is 39%.

Our next regression focuses on restaurants with significant amounts of ratings, in the 200–2,000 ratings range. The independent variable is now the average rating for each restaurant, and the dependent variable is the average rating of raters for that restaurant.

$$Y_j = a + b\,X_{2j}$$

$Y_j = \frac{1}{n_j}\sum_{\{users\ i\ who\ rated\ restaurant\ j\}} X_{1i}$ = average user rating, for users who rated restaurant j

where X_{1i} *and* X_{2j} are the average user rating and the average restaurant rating, respectively. Note that the index j ranges over restaurants rated between 200 and 2,000 times. Figure 3 displays a scatter plot of the relationship between the average restaurant rating and the average user rating, with the regression line.

The slope of the regression is 0.30, indicating that for each additional star a restaurant has, the average rating of their raters goes up by 0.30 stars. The R squared of this regression is 76%, indicating that a restaurant's score is significantly influenced by the type of raters that gave the rating, even for restaurants rated more than 200 times.

Are highly rated restaurants good at attracting inflating raters or are inflating raters good at finding highly rated restaurants? The average rating of inflating raters and deflating raters are 4.7 and 2.7, respectively. However, the average rating of restaurants rated by these two groups are 4.0 and 3.7, respectively. This regression to the mean indicates that the effect whereby inflating raters are better at choosing

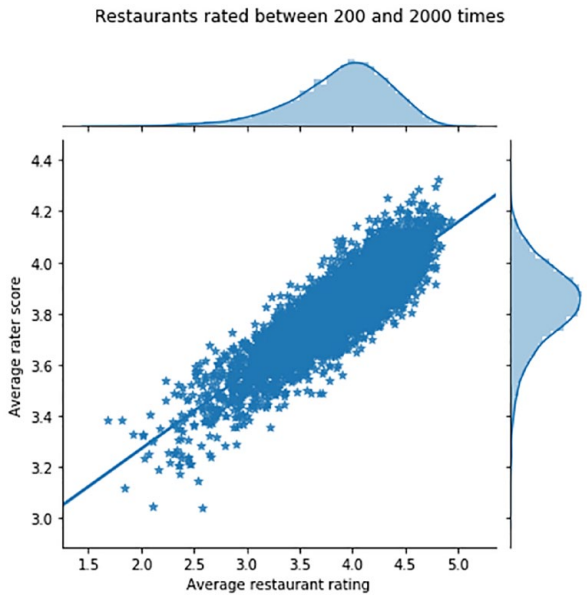

Figure 3.
A Scatter Plot of the Relation Between Average Restaurant Rating (Not Rounded) vs. Average Rater Score for That Restaurant.

good restaurants is relatively modest. The difference between the 4.0 and 3.7 averages may be partially explained by the inflating raters' own direct influence on the average rating.

Yelp Score Prediction

Figure 4 displays the main result of our article. The horizontal axis indicates the score categories (i.e., 3.5, 4.0, and 4.5), and the vertical axis represents accuracy scores. We have included a random baseline to compare the accuracy scores for deflating and inflating raters. Sampling deflating raters to sort the restaurants in our universe of 3,000 restaurants produces rankings that are better predictors of Yelp scores than inflating raters, for the top and bottom quartiles of restaurants. For instance, in the bottom quartile, the accuracy score for deflating raters is 56% (compared to 52% for inflating raters), and in the top quartile, the corresponding scores are 59% for deflating raters (compared to 49% for inflating raters). In the two central quartiles, the accuracy rates of both types of raters are similar and possibly not significant, as the error bars indicate. The numerical score coming from deflating raters is lower than that of inflating raters, which is why restaurants do not tend to like these users, much like the fear a restaurant may feel toward a hard-nosed Michelin critic. However, these results show that these users are more informative when it comes to producing reliable rankings.

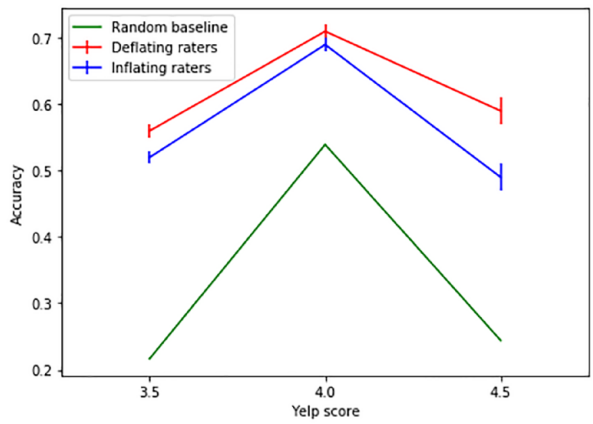

Figure 4.
Accuracy Scores for Random Baseline vs. Deflating and Inflating Raters.

Discussion

Implications for Theory

Our study primarily enriches existing research on strategies for mitigating the negative consequences of inflated ratings. Extant research has focused mostly on possible changes to the rating system, including changes in the ratings interface (Garg & Johari, 2019a; J. M. Kim et al., 2023; Leoni & Boto-García, 2023; Shahout et al., 2023). In contrast to these upstream approaches, our findings suggest that the impact of inflated ratings can be mitigated via a "post-processing" approach. By filtering raters and selecting a subset of deflating raters, the predictive accuracy of ratings, in particular at the top and bottom quartiles of the distribution, can be improved. This suggests that the data provided by deflating raters may be of higher value than those of inflating raters.

One possible limitation of our result is that taking the ratings of restaurants rated more than 200 times as a proxy for quality may be questioned. The strong correlation between these restaurant's scores and the average behavior of their raters (Figure 3) could just mean that restaurants with a 4.5 score are better at attracting the kind of raters who are generous. However, since there is significant trust in systems like Yelp, we can reasonably expect that restaurants with more than 200 ratings are closer to an idealized objective notion of quality.

Our results also touch indirectly on the economic impact of ratings. If ratings are a true reflection of restaurant quality, then we expect that in a fair environment, higher-rated restaurants will do better economically. Our study therefore highlights reasons why the link between ratings and economics is less than perfect. First, we find that the average rating of a restaurant can often be explained in terms of the

generosity or pickiness of the raters. Though this may be particularly harmful to restaurants with few ratings, we find that this phenomenon persists, even for restaurants rated more than 200 times. Second, we find that ratings inflation can weaken the link between ratings and the economic performance of a restaurant.

Managerial Implications

Our findings have a number of interesting implications for managerial practice. In the following section, we will discuss those from the perspective of ORS platforms and from the viewpoint of restaurant managers.

Implications for ORS Platforms. Our findings suggest that not all restaurant raters are created equal. In particular, we find that "deflating" raters, that is, those who display a tendency to be more critical in their average ratings of all the restaurants they have assessed, are better at predicting whether a particular restaurant belongs to the top or the bottom end of the distribution. In other words, as compared to the raters that are of the "inflating" kind, "deflating" raters seem to be more discerning experts that are better at identifying the "true gems" and the "duds," two categories of restaurants that users of ORS platforms particularly care about.

For ORS platforms, this would provide the opportunity to compute a differentiated restaurant score based exclusively on the ratings of the more critical "deflating" raters. Such a differentiated scoring system would bear some similarity with the "audience score" vs. the "critics score" on Rotten Tomatoes, a popular review-aggregation website for film and television. While the current Yelp score would reflect the more inflating "audience" score, with ratings that are generally biased upward and condensed at the upper end of the rating scale, an additional "Yelp critics" score based on the ratings of "deflating" raters would have more in common with the "critics" score which reflects the opinion of more discerning expert judges.

From a guest perspective, such a new score would arguably provide much richer information. The current problem with ORS scores is a general tendency toward grade inflation, leading more and more users to rely exclusively on the content of the comments, rather than the overall numerical rating. A new score based on critical raters would have greater diagnostic value in and by itself. In particular, it would help guests single out promising restaurants that are at the top of the distribution and avoid those that are at the lower end.

Implications From the Restaurant Perspective. For restaurateurs, a "Yelp critics" score could also be beneficial. While one could argue that restaurants care above all for high average ratings, they also suffer from the fact that grade inflation reduces an individual restaurant's ability to stand out from the crowd. A high score on "Yelp critics" would therefore be a more potent marketing argument than the current scoring system.

In addition, if restaurants would receive information not only about their average score but also about the average scores of the raters that assessed them, this would allow for a prediction of the direction in which their score is likely to evolve in the future. As Figure 3 suggests, average restaurant ratings and average rater scores are highly correlated. Therefore, if a restaurant knows that the average scores of its raters are higher than its average score, regression to the mean would suggest that in the future, the average rating of this restaurant will go down, as over time, a higher proportion of critical raters will weigh in. Conversely, if the average scores of its raters are lower than its average score, an improvement of the average restaurant score is to be expected over time.

The latest argument is particularly important for restaurants with limited numbers of ratings. When such restaurants, that may have been opened recently, have received a low rating and the raters are particularly picky, their scores are likely to pick up substantially as more raters are added.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, or publication of this article.

ORCID iD

Steffen Raub https://orcid.org/0000-0003-2092-0704

References

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.

Athey, S., Castillo, J. C. & Chandar, B. (2019). Service quality in the gig economy: Empirical evidence about driving quality at Uber, Available at SSRN.

Aziz, A., Li, H., & Telang, R. (2023). The consequences of rating inflation on platforms: Evidence from a quasi-experiment. *Information Systems Research*, *34*(2), 590–608.

Dey, D., & Kumar, P. (2023). Do psychological attributes of online review play role in predicting rating? An empirical investigation. *Computers in Human Behavior*, *148*, 107895.

Filieri, R., Raguseo, E., & Vitari, C. (2019). What moderates the influence of extremely negative ratings? The role of review and reviewer characteristics. *International Journal of Hospitality Management*, *77*, 333–341.

Filippas, A, Horton, J.J., & Golden, J.M. (2022) Reputation inflation. *Marketing Science*, *41*(4), 733–745.

Gao, G., Greenwood, B. N., Agarwal, R., & McCullough, J. (2015). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality. *MIS Quarterly*, *39*(3), 565–590.

Garg, N., & Johari, R. (2019a). *Designing informative rating systems: Evidence from an online labor market*. http://arxiv.org/abs/1810.13028

Garg, N., & Johari, R. (2019b). *Designing optimal binary rating systems*. https://arxiv.org/abs/1806.06908

Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, *61*(2), 142–153.

Hlee, S., Lee, J., Yang, S.-B., & Koo, C. (2019). The moderating effect of restaurant type on hedonic versus utilitarian review evaluations. *International Journal of Hospitality Management*, *77*, 195–206.

Horton, J. J., & Golden, J. M. (2015). *Reputation inflation in an online marketplace [Working paper]*. New York University Stern School of Business.

Hu, N., Pavlou, P. A., & Zhang, J. (2017). On self-selection biases in online product reviews. *MIS Quarterly*, *41*(2), 449–471.

Jiang, Y., & Guo, H. (2015). Design of consumer review systems and product pricing. *Information Systems Research*, *26*(4), 714–730.

Kim, J. M., Liu, J., & Yousaf, S. (2023). Does change in the scoring system impact service evaluation? Evidence from Booking.com. *International Journal of Contemporary Hospitality Management*, *35*(3), 828–847.

Kim, W. G., Li, J., & Brymer, R. A. (2016). The impact of social media reviews on restaurant performance: The moderating role of excellence certificate. *International Journal of Hospitality Management*, *55*, 41–51.

Kwok L., Xie K. (2016). Factors contributing to online review helpfulness: Do manager responses play a role? *International Journal of Contemporary Hospitality Management*, *28*, 2156–2177.

Lee, H., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing*, *28*(7), 675–688.

Lee, M., Kwon, W., & Back, K.-J. (2021). Artificial intelligence for hospitality big data analytics: Developing a prediction model of restaurant review helpfulness for customer decision-making. *International Journal of Contemporary Hospitality Management*, *33*(6), 2117–2136.

Leoni, V., & Boto-García, D. (2023). "Apparent" and actual hotel scores under Booking.com new reviewing system. *International Journal of Hospitality Management*, *111*, 103493.

Li, H., Liu, H., & Zhang, Z. (2020). Online persuasion of review emotional intensity: A text mining analysis of restaurant reviews. *International Journal of Hospitality Management*, *89*, 102558.

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

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140–151.

Luca, M. (2016). *Reviews, reputation, and revenue: The case of Yelp.com* (Working paper 12-016). SSRN Electronic Journal. http://www.ssrn.com/abstract=1928601

Luca, M., & Zervas, G. (2015). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412–3427.

Luo, Y., Tang, L., Kim, E., & Wang, X. (2020). Finding the reviews on Yelp that actually matter to me: Innovative approach of improving recommender systems. *International Journal of Hospitality Management*, *91*, 102697.

Ma, Y., Xiang, Z., Du, Q., & Fan, W. (2018). Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *International Journal of Hospitality Management*, *71*, 120–131.

Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly*, *34*(1), 185–200.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, *50*, 67–83.

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

Salehi-Esfahani, S., & Ozturk, A. B. (2018). Negative reviews: Formation, spread, and halt of opportunistic behavior. *International Journal of Hospitality Management*, *74*, 138–146.

Sayfuddin, A., & Chen, Y. (2021). The signaling and reputational effects of customer ratings on hotel revenues: Evidence from TripAdvisor. *International Journal of Hospitality Management*, *99*, 103065.

Shahout, R., Peisakhovsky, Y., Stoikov, S., & Garg, N. (2023). *Interface design to mitigate inflation in recommender systems*. http://arxiv.org/abs/2307.12424

Stamolampros, P., & Korfiatis, N. (2018). Exploring the behavioral drivers of review valence: The direct and indirect effects of multiple psychological distances. *International Journal of Contemporary Hospitality Management*, *30*(10), 3083–3099.

Torres, E. N., Singh, D., & Robertson-Ring, A. (2015). Consumer reviews and the creation of Booking transaction value: Lessons from the hotel industry. *International Journal of Hospitality Management*, *50*, 77–83.

Wang, E. Y., Nang Fong, L. H., & Law, R. (2022). Detecting fake hospitality reviews through the interplay of emotional cues, cognitive cues and review valence. *International Journal of Contemporary Hospitality Management*, *34*(1), 184–200.

Wang, F. X., & Anderson, C. (2023). How firm strategies affect consumer biases in online reviews. *Service Science*, *15*(3), 172–187.

Wang, X., Tang, L., & Kim, E. (2019). More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? *International Journal of Hospitality Management*, *77*, 438–447.

Wang, Y., Kim, J., & Kim, J. (2021). The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity. *International Journal of Hospitality Management*, *95*, 102895.

Yang, S.-B., Shin, S.-H., Joun, Y., & Koo, C. (2017). Exploring the comparative importance of online hotel reviews' heuristic attributes in review helpfulness: A conjoint analysis approach. *Journal of Travel & Tourism Marketing*, *34*, 963–985.

Yoo, T., & Suh, K.-H. (2022). Experts vs. the public in the evaluation of restaurants: A business ecosystem approach. *International Journal of Hospitality Management*, *105*, 103265.

Zervas, G., Proserpio, D., & Byers, J. W. (2017). A first look at online reputation on Airbnb, where every stay is above average. *Marketing Letters*, *32*, 1–16.

Zhang, Z., Li, Y., Li, H., & Zhang, Z. (2022). Restaurants' motivations to solicit fake reviews: A competition perspective. *International Journal of Hospitality Management*, *107*, 103337.

Zhu, D. H., Zhang, Z. J., Chang, Y. P., & Liang, S. (2019). Good discounts earn good reviews in return? Effects of price promotion on online restaurant reviews. *International Journal of Hospitality Management*, *77*, 178–186.

Author Biographies

Sasha Stoikov is a Senior Research Associate at Cornell Financial Engineering Manhattan. He holds a B.S. from MIT and a M.S. from the University of Wisconsin, Madison, in addition to his Ph.D. from the University of Texas. He has worked as a VP in the High Frequency Trading group at Cantor Fitzgerald. Sasha's research is in market microstructure, dynamics of limit order books, market-making techniques and music recommendation algorithms.

Stefano Borzillo is an associate professor at EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland. Stefano obtained his Ph.D. in Organizational Behavior from the University of Geneva and was a Visiting Research Fellow at the Stern School of Business (New York University). His main fields of research are knowledge networks, communities of practice, processes of knowledge creation, best practice transfers in organizations, and talent management systems.

Dr. Steffen Raub, is Full Professor of Organizational Behavior at EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland. He holds a PhD from the University of Geneva, Switzerland, and is an Editorial Board member of the Cornell Hospitality Quarterly. His research focuses on empowerment, organizational citizenship behavior, proactive service behaviors and sustainability/corporate social responsibility.