Platform-Generated Quality Ratings: Theory, Empirics and Welfare Implications

Jie Bai, Yanyou Chen, Daniel Yi Xu, Zhe Yuan*

PRELIMINARY DO NOT CITE

This Version: *October 23, 2023*

Abstract

In order to address the issue of asymmetric information and promote market regulation, digital platforms have implemented platform-generated rating (PGR) systems to provide quality information to the market. Unlike user-generated rating (UGR) systems, PGRs rely on platform-generated data including product quality, customer service, and logistics. To better understand the impact of PGRs, we conducted a study in collaboration with a large E-commerce platform, exploring both empirical and theoretical implications on consumer beliefs, sellers' quality incentives, and market outcomes. Our empirical analysis utilized a regression-discontinuity method to examine customer response to high PGR sellers. Results indicate that customers tend to purchase more from high PGR sellers, suggesting that PGRs can be effective in signaling quality to consumers. We also found that the presence of a PGR system affects sellers' quality incentives and the distribution of equilibrium quality. These findings demonstrate that PGRs have the potential to promote quality competition among sellers in the market. Finally, we conducted a counterfactual experiment to evaluate the welfare consequences of PGR adoption. Our (tentative) results suggest that PGR adoption can have a positive impact on overall welfare. Overall, our study highlights the potential benefits of PGRs in promoting market efficiency and enhancing consumer welfare.

Keywords: Two-Sided Market, Platform Governance, Optimal Rating

*Contact: Bai: Harvard Kennedy School, jie_bai@hks.harvard.edu; Chen: University of Toronto, yanyou.chen@utoronto.ca; Xu: Duke University and NBER, daniel.xu@duke.edu; Yuan: Zhejiang University, yyyuanzhe@gmail.com.

1 Introduction

E-commerce sales have grown tremendously in recent years, reaching $4.9 trillion and 19% of global retail sales in 2021. The rapid expansion of the online marketplace presents new growth opportunities to firms by lowering the entry barrier and allowing many firms to sell their products online. However, the very same feature also weakens the market selection mechanism by attracting a large number of sellers with heterogeneous quality to compete in the online marketplace. Consumers cannot perfectly assess the true quality of a given seller and rely on consumer-generated reviews, which are only noisy signals of quality (Tadelis, 2016; Hui et al., 2020; Li et al., 2020). Existing studies have shown that substantial information friction and misallocation remain unresolved in the online marketplace despite the presence of existing search ranking and review mechanisms (Bai et al., 2022). Such information friction may further distort sellers' incentives to invest in high quality.

Increasingly, platforms have started to design and introduce platform-generated rating (PGR) systems to inject quality information to the market. Different from traditional consumer or user-generated ratings (UGR), PGRs are based on objective measures of quality that are directly tracked by the platform, through data generated by actual transactions and interactions between the two sides of the market. Many of these objective metrics may not be observed or only imperfectly observed by consumers. Therefore, introducing a PGR gives consumers additional information about a seller's underlying quality, which can help to resolve informational uncertainty and improve market outcomes. Furthermore, by designing a PGR, the platform can better control and guide sellers' effort incentives in investing in high quality (Vatter, 2022). In practice, PGRs are commonly based on a seller's relative performance amongst a large pool of market participants, as opposed to following some absolute quality standards. Theoretically, while relative performance-based rating insures sellers against common shocks (Gibbons and Murphy, 1990), thereby allowing the platform to extract more surplus from them, its implications on quality incentives, pricing and the resulting consumer welfare are far less clear. Empirically, there is a lack of rigorous systematic analysis of PGRs on both firm and market outcomes.

Our research aims to empirically and theoretically analyze the impacts of platform-generated ratings on consumer beliefs, sellers' quality incentives and market outcomes. We address three questions. First, what is the additional value of platform generated information in a mature market? Do consumers respond to it? Second, how does the platform information disclosure affect supplier's effort choice? Third, does the current ranking mechanism help to avoid the shirking of suppliers with strong cumulative track record?

We collaborate with one of the largest online retail platforms in the world, which hosts millions of sellers and features billions of products. The platform introduced a relative performance rating system based on objective quality measures along the following five dimensions: (1) product quality (measured by the refund rate and consumer ratings); (2) pre-sales service quality (measured by customer service response speed/rate); (3) logistic speed (measured by the delivery time); (4) post-sales refund service quality (measured by the refund speed); (5) dispute rate. Sellers are first ranked based on their performance in each metric and then are assigned a score ranging from 3 to 5 in incremental of 0.5 based on their rank. For instance, sellers in the top quantile of delivery time receive a score of 4.5 or 5, while the bottom quantile receive a score of 3 or 3.5. After that, the platform calculates a weighted-average score by aggregating the scores across individual metrics, where the weights are disclosed to all market participants.

The objective quality metrics and the scores are updated every two weeks, which helps to sustain the incentive to maintain high quality among sellers. Comparing the PGR with the existing UGR, the former potentially covers a broader set of quality dimensions and its distribution is much more dispersed: by design, only one-fifth of sellers could receive a PGR above a score of 4, whereas UGR is generally much more skewed towards the top and less differentiating as shown in the previous literature (Tadelis, 2016).

We have assembled a rich micro dataset containing detailed high-frequency objective quality measures for a seller. We also have full information about the platform's exact formula to aggregate the individual quality measures into a relative performance score. We complement the quality data with order transactions as well as consumer clicks and purchases to shed light on the demand-side responses to information. For the empirical analysis, we have obtained proprietary data of a random sample of 255,503 sellers and their transactions from Apr. 2020 to Mar. 2021. The data contains detailed high-frequency measures of a seller's objective performance along a wide spectrum of quality dimensions, as well as sales performance measured in terms of the number of daily visits, number of orders, number of purchase payments and their amounts. We further compute the conversion rate (CVR), which is the number of orders divided by the number of visits. As mentioned above, we also have full information about the exact formula that the platform uses to aggregate the multi-dimensional quality measures and the percentile cutoffs used to convert each seller's quality performance into a relative performance score ranging from 3 to 5.

Our analysis will proceed in three steps. First, we perform reduced-form analysis to examine how consumers respond to changes in the PGR of a seller. Similar to a regression discontinuity framework, we leverage the high-frequency panel data and explore changes in the PGR and sales over very short time windows (consecutive days before and after each rating update). Second, we build a model that features heterogeneous sellers with differential costs of investing in quality and imperfect mapping from quality investment effort to actual quality realization due to both common and idiosyncratic shocks. We theoretically examine how quality incentives and equilibrium quality distribution are affected by a PGR that is based on relative performance. Finally, we structurally estimate the model to quantify the welfare effects of introducing a PGR and use the model to evaluate alternative designs of platform rating.

Our empirical findings show that one star increase in rating results in 0.27% increase in the number of visitors, 0.89% increase in the number of buyers (visitors who make a purchase), 1.55% increase in the transaction amount and seller CVR increase by 0.08 percentage point.

2 Institutional Details

We collaborate with one of the world's largest business-to-consumer online retail platforms. The platform mandates that sellers must be registered companies, including brand owners or authorized distributors. With millions of active monthly users, the platform hosts billions of product listings, and customers can interact with millions of sellers.

The platform captures comprehensive data on the everyday operations of the sellers, such as the quality of customer service, customer ratings and feedback on product quality, delivery speed, and the resolution of any disputes. The vast amount of data generated on these operations, commonly referred to as big data, enables the platform to design and generate ratings that provide valuable insights for customers and sellers.

2.1 Platform-generated Rating

In July 2018, the platform introduced the Platform-generated Rating system (PGR) in addition to the classic user-generated Seller Ratings where customers can rate sellers (UGR). The PGR is a star-level rating system that assigns a star level ranging from 3 to 5 stars in incremental steps of 0.5 based on the relative service quality of sellers. The PGR star level is based on the seller's performance in five quality dimensions: product quality, consult service, logistic speed, refund speed, and dispute rate. The PGR system uses an algorithm to calculate the seller's overall performance in each of these dimensions.[1] It is important to note that the PGR's star level is relative and based on the performance of other sellers on the platform. Only the top percentile of sellers can achieve a score of 4.5 or 5, which reflects exceptional performance.

[1] See Appendix A for details.

The PGR serves as an additional tool for customers to assess seller performance beyond the classic UGR rating. The PGR star level is prominently displayed both at the top of the seller store front page and in the middle of the product description page (Figure 1). Every two weeks, the platform updates all sellers' PGR star level and reveal them to the public.[2] Sellers can track their performance in each dimension as well as the over all performance through a data module.

Figure 1: PGR display

To sum up, the platform-generated ratings provide a reliable measure of seller performance that is based on extensive customer subjective feedback and objective performance metrics. These ratings are generated using algorithms that integrate various aspects of seller performance, including pre-sale, during-sale, and post-sale service quality. By assigning a relative score, the PGR incentivizes sellers to continuously improve their service quality to maintain or improve their rating and help customers make informed purchase decisions.

2.2 The purpose of PGR

The introduction of the Platform-generated rating (PGR) by the platform has emerged as a promising solution to address the challenge of asymmetric information and improve seller quality regulation. Customers often face the problem of insufficient information about seller quality (Akerlof, 1970), which can significantly impact their purchase decisions. Reputation systems, such as Detailed Seller Ratings (a type of user-generated rating, UGR), are designed to overcome this issue. However, UGR has certain limitations. Firstly, only customers who complete the purchase can provide a UGR, while unsatisfied buyers who do not make a purchase cannot submit a rating. Secondly, some dissatisfied buyers may leave the platform without leaving any feedback, further exacerbating the issue (Tadelis, 2016). Lastly, UGR is a subjective measure that can be manipulated by buyers.

PGR offers a promising approach to address the limitations of UGR. It includes five dimensions, consisting of around ten indices, with UGRs being a subset of the indices. Additionally, PGR measures pre-sale, during-sale, and post-sale service quality, including product quality, consult service, logistic speed, refund speed, and dispute rate. Most of the PGR indices are objective and based on algorithms, providing a more accurate and reliable measure that is difficult to manipulate.

[2]The platform updates the PGR every 1st and 16th day of the month.

On the seller side, PGR can incentivize investment in various quality dimensions to improve the quality ranking and attract more customers to the store. As sellers expect that higher PGR scores can increase conversion rates and attract more customers (two-sided market), they have a natural incentive to invest in the quality dimensions measured by PGR.

In conclusion, the platform's introduction of PGR has provided a more comprehensive and reliable measure of seller quality, addressing the limitations of existing reputation systems such as UGR. Furthermore, it has incentivized sellers to invest in various quality dimensions to improve their rankings, leading to higher conversion rates and increased customer trust.

3 Data

3.1 Data Description

We obtained proprietary data from a sample of sellers on the platform for the period of April 2020 to March 2021. The data includes the number of daily visitors, number of daily paying customers, number of daily orders, and daily transaction amount for each sellerwhich allow us to calculate the conversion rate (CVR) as orders divided by visitors, as well as scores for different quality dimensions, overall quality score, and the UGR and PGR star level.

To match the bi-weekly updating cycle of the PGR star level, our database is organized at the seller-bi-week level, with two observations per month for each seller.[3] Each observation records the PGR star level observable to the customers throughout the period, different quality dimensions, and the customer visit, purchase and conversion metrics before and after the star level changes.

We focused on a select sample of sellers across the top 13 industries, and only included observations with consecutive operations.[4] This means that sellers must have had at least five paying customers every bi-week for the past three months to be included in our sample. Excluding sellers without consecutive operations ensures that demand and transactions are stable, and that the PGR star level accurately reflects their quality level.

3.2 Summary Statistics

Table 1 reports the summary statistics of the data. Our sampled dataset contains 255,503 sellers and 4.6 million observations.[5] Given that our cooperating platform is a business-to-

[3]Specifically, for each seller in each month, there are two observations. The first observation covers from the first day to the 15th day of the month. The second observation covers from the 16th day of the month to the end of the month.

[4]The top 13 industries contribute to over 85% platform transactions.

[5]Note that our dataset is not a balanced panel because we only keep consecutive operations.

customer platform, the sellers have a mean bi-weekly transaction of $441,047$ CNY, which is equivalent to $\$63,000$.[6]

Table 1: Data Summary

	Mean	Std	25 Percentile	Median	75 Percentile
# Visitors	81,742	755,548	4,984	16,347	52,144
# Buyers	3,673	61,093	145	525	1,839
GMV	441,047	15,841,434	20,192	66,489	206,159
PGR Star Level	3.91	0.41	3.5	4	4
PGR Score Value	3.75	0.48	3.45	3.79	4.08
# Sellers			255,503		
# Obs			4,601,757		

Table 2 reports the transition probability across periods.[7] A seller is more likely to maintain in its current star level. For instance, a four-star seller is 72.6% likely to be a four star seller in the next period. It is rare to observe a seller star change dramatically. For instance, a four-star seller is only 0.1% chance to become a three-star seller in the next period and 2.2% probability to be a five-star seller.

Table 2: PGR Star Level Transition Probability

	PGR Star Level (next)				
PGR Star Level	3	3.5	4	4.5	5
3	41.5%	54.4%	4.0%	0.1%	0.0%
3.5	2.2%	72.0%	24.8%	0.9%	0.2%
4	0.1%	18.0%	72.6%	7.1%	2.2%
4.5	0.0%	2.5%	32.7%	63.0%	1.8%
5	0.0%	1.3%	24.0%	3.2%	71.5%

3.3 Compare PGR with UGR

We conducted a comparative analysis of seller URG and PGR levels, and the findings are presented in Table 6. The table provides a summary of the PGR star level and UGR levels, allowing for an evaluation of the relative quality of the sellers. In addition to the PGR star level, we also calculated an "Absolute PGR" index, which serves as an absolute measure of seller quality. To obtain the Absolute PGR, we employed a regression analysis with the PGR star as the dependent variable and all quality measures as independent variables. The resulting fitted value represents an absolute measure of seller quality, as all quality measures

[6] We use an exchange rate of 7 CNY to 1 USD.

[7] The transition matrix measured by the number of sellers are reported in Appendix B.

are expressed as absolute values instead of relative performance of the sellers. By utilizing the Absolute PGR index, we can more accurately evaluate the quality of the sellers.

Our analysis indicates that the PGR metric exhibits greater variability than the UGR metric. Specifically, PGR ranges from three to five with a standard deviation of 0.48, whereas the UGR metric has an average score of 4.89, with a standard deviation of 0.07, despite its potential range of 1 to 5. This is largely attributed to the fact that very few customers leave negative comments, resulting in a higher average score for UGR. The greater variability of the PGR metric compared to the UGR metric may be advantageous for customers in distinguishing high quality sellers from low quality ones.

Table 3: PGR and UGR summary

VARIABLES	Mean	Std. Dev.	5 pcl	25 pcl	Median	75 pcl	95 pcl
PGR Score Value	3.75	0.48	2.91	3.47	3.80	4.09	4.45
Absolute PGR Score Value	3.75	0.32	3.17	3.63	3.82	3.95	4.11
UGR (description)	4.89	0.07	4.77	4.86	4.90	4.94	4.99
UGR (logistics)	4.89	0.07	4.77	4.85	4.90	4.94	4.98
UGR (service)	4.88	0.08	4.75	4.84	4.89	4.93	4.98

The PGR and UGR metrics share some commonalities while also exhibiting differences in their computation and application. Certain UGR measures, such as UGR (description) and UGR (logistics), are encompassed within the PGR, albeit with different periodicity in their computation. UGR displayed on the platform is based on customer ratings in the past 180 days, whereas PGR is derived using customer ratings in the past 30 days. It is worth noting that a positive correlation exists between customer ratings (UGR) and seller service quality (measured by PGR), indicating the potential interdependence of these metrics. To delve deeper into the correlation between PGR and three UGR indices, we conducted an analysis, the results of which are presented in Table 4. Our findings reveal a positive relationship between PGR and UGR, with correlation coefficients ranging around 0.16. Furthermore, we observed a high correlation among the UGR indices, typically exceeding 0.7. These insights can assist businesses in comprehending the relationship between seller service quality and customer ratings, informing the development of effective strategies to enhance customer satisfaction and drive business growth.

Table 5 presents a summary of the distinctions between PGR and UGR. While PGR encompasses nine quality measures, UGR covers only three indices. Specifically, PGR takes into account quality measures across various stages of the sales process, including pre-sale, during-sale, and post-sale. Conversely, only customers who complete a transaction can provide a UGR. Additionally, PGR exhibits a larger variation, ranging between three to five stars, while UGR demonstrates a smaller range, with most ratings being above 4.9

Table 4: Correlation between PGR and UGR

	PGR	UGR(product)	UGR(logistic)	UGR(service)
PGR	1			
UGR(product)	0.1555	1		
UGR(logistic)	0.1596	0.6929	1	
UGR(service)	0.1675	0.7926	0.862	1

out of 5. Furthermore, it is more challenging for sellers to manipulate PGR than UGR because of the broader spectrum of quality dimensions involved, some of which are difficult to falsify. For instance, a seller can potentially fake a UGR by posing as a customer and leaving a comment. However, it is challenging to fabricate a logistic record, which is one of the components included in PGR. Moreover, investing in logistics or customer service can boost a seller's PGR within two weeks. Conversely, it may take several months for UGRs to reflect changes made by the seller. Finally, platform designs for PGR can be easily modified by introducing new indices or changing the weight or ranking of each component. However, altering UGR designs is relatively more challenging for platforms to accomplish.

Table 5: Comparision between PGR and UGR

Dimensions	PGR	UGR
Variation	Large	Small
Number of Indices	Aggregate Nine sub-indices to one	Three
Coverage	pre-/during/post-sale	Only Post-sale
Manipulatable	Relatively Difficult	Easy (fake reviews)
Seller Feedback	Fast and Direct	Slow and Indirect
Platform Design	Easy	Difficult

Table 6 displays the distribution of PGR scores and UGRs by PGR star level. The results demonstrate that PGR star levels can effectively differentiate seller qualities. Specifically, an increase in PGR star level is associated with a consistent and monotonic increase in all UGR metrics.

Table 6: The distribution of PGR scores and UGR Scores across PGR star level

Star Level	3	3.5	4	4.5	5
PGR Score Value	2.21	3.31	3.93	4.31	4.35
Absolute PGR Score Value	2.84	3.54	3.85	4	4.02
UGR (description)	4.84	4.88	4.9	4.91	4.92
UGR (logistics)	4.84	4.88	4.89	4.9	4.91
UGR (service)	4.82	4.87	4.89	4.9	4.91

Table 7 presents the distribution of PGR scores for different UGR levels. Notably, even among sellers with a UGR of five, indicating a high level of customer satisfaction, there is still a significant dispersion of PGR scores. Specifically, approximately 31% of these sellers have a PGR of 3.5 or lower, indicating a relatively low quality of service. Such sellers are more likely to be small-scale and receive few orders, and are fortunate to have received few negative ratings.

Table 7: Relationship between UGR, PGR and UGR Score

	PGR				
UGR (mean)	3	3.5	4	4.5	5
5	1%	30%	48%	15%	6%
4.95-4.99	1%	31%	51%	12%	6%
4.9-4.94	1%	30%	50%	13%	6%
<4.9	2%	42%	46%	8%	3%

4 Reduced-Form Analysis

We conduct a reduced-form analysis to investigate how consumers react to changes in the platform-generated-rating (PGR) of a seller. Our objective is to examine whether changes in a seller's PGR have a significant effect on sales outcomes. To do this, we utilize high-frequency panel data and examine the changes in quality ratings and sales over a span of two consecutive days. We focus on the immediate time frame just before and after each round of rating update, specifically within a one-day range. By doing so, we isolate the impact of consumers' expectations regarding the underlying service quality, as the updated ratings provide new information. This approach, similar to a regression discontinuity design, enables us to control for any unobservable actions taken by sellers that could potentially influence both quality ratings and sales over time. This is due to the fact that a seller's rating is determined by their performance across five quality dimensions over a two-week period. Consequently, it is not possible for sellers to manipulate their rating within a single day.

To examine the relationship between changes in sales outcomes and changes in the PGR, we regress changes in sales outcomes of a seller on changes in the PGR before and after each rating update, controlling for both seller and time fixed effect of each update:

$$\Delta Y_{jt} = \alpha + \lambda_t + \nu_j + \beta \cdot \Delta r_{jt} + \epsilon_{jt} \tag{1}$$

In this equation, Y_{jt} represents the sales outcome, including metrics such as the number of visitors, purchases, sales revenue, and conversion rate. The terms λ_t and ν_j capture the fixed effects for time and sellers, respectively. Δr_{jt} represents the change in PGR for seller j. The parameter β captures the combined influence of belief updating and the willingness of buyers to pay for higher quality. More specifically, we can express it as:

$$\beta \cdot \Delta r_{jt} = \beta_r \cdot (E[\xi_{jt}|r_{jt}] - E[\xi_{jt}|r_{jt-1}]) \tag{2}$$

Here, β_r measures the willingness of buyers to pay for service quality, while $E[\xi_{jt}|r_{jt}]$ represents the expected value of the underlying service quality given the rating of seller j at time t.

The regression results are presented in Table 8. Columns (1) to (4) display the outcomes for the number of visitors, number of buyers, transaction amount (GMV), and conversion rate respectively. The findings indicate a significant impact of changes in the PGR on sales outcomes. Specifically, a one-star increase in the rating leads to a 0.27% rise in the number of visitors, a 0.89% increase in the number of buyers (i.e., visitors who make a purchase), a 1.55% increase in transaction amount, and an improvement of 0.08 percentage points in the purchase conversion rate. Remarkably, these estimated effects are quite large, particularly when considering that our analysis focuses on changes observed over a mere one-day timeframe. This highlights the considerable impact of changes in the PGR on consumer behavior, revealing their strong responsiveness to such variations.[8]

Table 8: Demand Estimation: Regression Discontinuity (One Day Difference)

	(1)	(2)	(3)	(4)
	Δln(# of Visitors)	Δln(# of Buyers)	Δln(GMV)	ΔCVR
ΔStar	0.00267***	0.00890***	0.0155***	0.000825***
	(0.000402)	(0.000920)	(0.00249)	(0.000173)
Constant	-0.0151***	0.0783***	0.110***	0.00256***
	(0.000603)	(0.00140)	(0.00314)	(0.0000714)
# Obs	4,601,744	4,601,744	4,601,744	4,600,836
R-squared	0.055	0.264	0.083	0.002

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Subsequently, we investigate the heterogeneous impact of changes in the PGR on different sellers. Specifically, we categorize sellers into 10 tiers based on their GMV, with tier

[8] As part of our robustness check in Appendix D, we explore alternative timeframes by extending the analysis from one-day to three- or five-day periods. The corresponding results can be found in Tables D.1 and D.2, which are consistent with those reported in Table 8.

1 representing the largest sellers and tier 10 representing the smallest. The findings are presented in Figure 2.

Our analysis reveals that small sellers (those with lower GMV) experience a higher increase in the number of visitors when their star levels rise. Conversely, large sellers (those with higher GMV) experience larger increase in the number of purchases, higher GMV, and an improved conversion rate following an increase in their star levels. These results indicate that while an increase in the PGR leads to a rise in visitor traffic for small sellers, it does not necessarily translate into a corresponding increase in sales. On the other hand, for large sellers, the increase in PGR has a positive impact on both conversion and overall sales performance.

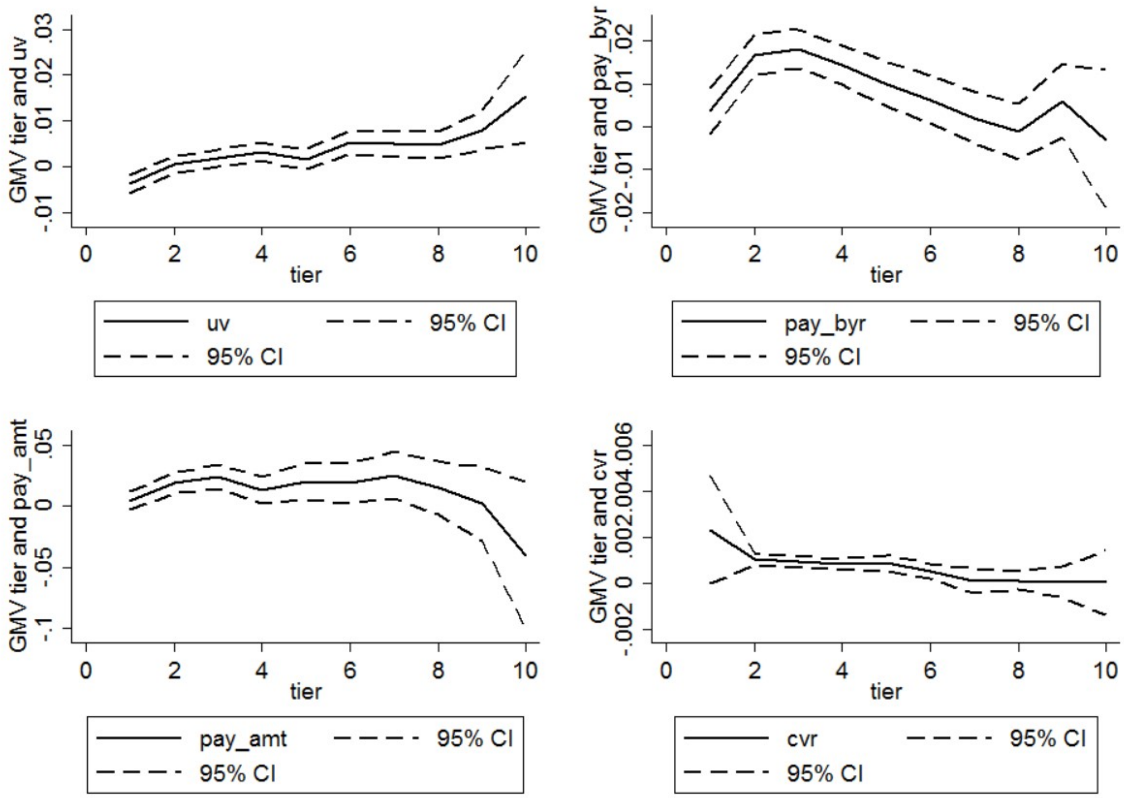

Figure 2: Demand Estimation by GMV Tier

5 Model

The market consists of N sellers, where each seller j draws an ex ante cost of quality investment c_j from the distribution $F_c(\cdot)$. The sellers then select their quality investment level, and their ex post quality ξ_j determines their ranking by the platform. The top τ percentile sellers are labeled as H-type, and the remaining sellers are labeled as L-type. The underlying distribution $F_c(\cdot)$ is common knowledge to both sellers and consumers. Consumers only know about the type of each seller but not the underlying quality, and use this information to choose which seller to purchase from.

The timing assumption is that the platform announces the rating mechanism first, then sellers make investment decisions, and after the ex post quality is realized, sellers make pricing decisions. Consumers then make their demand choices based on the ratings and prices of each seller. We solve the model backwards.

Rating mechanism τ	Quality investment e_j	Realized quality q_j	Prices p_j	Demand D_j	
1	2	3	4	5	t
Platform	**Sellers**	**Nature**	**Sellers**	**Consumers**	

5.1 Demand Model

Consumers choose a specific seller or an outside option. Each seller has a seller (or seller-quarter) fixed effect that captures the breadth of products (characteristics and prices) and long-term brand name. Consumers also value the hidden "service quality" provided by a seller, which can only be inferred from the platform ratings. After forming expectations regarding the ex post quality of each seller based on their observed type, consumers make a discrete choice regarding which seller to purchase from.

The utility that consumer i receives from seller j is given by the following equation:

$$\begin{aligned} u_{ij} &= \alpha p_j + \boldsymbol{X_j \beta} + \beta_r \cdot E[\xi_j | r_j] + \epsilon_{ij} \\ &= \delta_j + \beta_r \cdot E[\xi_j | r_j] + \epsilon_{ij} \end{aligned} \tag{3}$$

Here, p_j represents the price of seller j, $\boldsymbol{X}_j$ is a vector of characteristics specific to the seller, β_r is the consumer's preference for quality, and $E[\xi_j|r_j]$ is the expected quality of seller j given its rating r_j. The variable ϵ_{ij} represents a type of error that follows an extreme-value distribution and is not observable to the econometrician.

Under the assumption of a standard Type-I Extreme Value distribution, we can calculate the demand of seller j as

$$D_j = M \cdot \frac{\exp(\delta_j + \beta_r \cdot E[\xi_j | r_j])}{D_0 + \sum_{j'} \exp(\delta_{j'} + \beta_r \cdot E[\xi_{j'} | r_{j'}])}, \tag{4}$$

Here, M represents the mass of demand, and D_0 represents the demand for outside options.

We assume that consumers are rational and **have complete information**. They are aware of the underlying distribution of quality investment costs for sellers, which follows $c_j \sim F_c(\cdot)$, and they are also aware of the optimal level of investment that sellers will make, conditional on c_j. For sellers who are ranked as H-type, the expected quality is derived as:

$$E[\xi_j|H] = \int_{F_q^{-1}(\tau)}^{\infty} x \cdot \frac{f_q(x)}{1 - F_q(F_q^{-1}(\tau))} dx = \int_{F_q^{-1}(\tau)}^{\infty} x \cdot \frac{f_q(x)}{1 - \tau} dx \tag{5}$$

Here, $F_q(\cdot)$ represents the ex post quality distribution in equilibrium, and $F_q^{-1}(\tau)$ is the threshold quality level required to be ranked as an H-type seller, given the threshold τ.

The expected quality for sellers who are ranked as L-type is calculated as:

$$E[\xi_j|L] = \int_{0}^{F_q^{-1}(\tau)} x \cdot \frac{f_q(x)}{F_q(F_q^{-1}(\tau))} dx = \int_{0}^{F_q^{-1}(\tau)} x \cdot \frac{f_q(x)}{\tau} dx. \tag{6}$$

5.2 Seller's Problem

Sellers decide on their quality investment level and set their prices. During the product market competition stage, sellers optimize their pricing strategy to maximize their profits, which can be expressed as follows:

$$\max_{p_j} D_j(p_j, \boldsymbol{p_{-j}}, r_j, \boldsymbol{r_{-j}}) \cdot (p_j - w)$$

Here, D_j represents the market demand for store j given its price p_j, prices of other stores $\boldsymbol{p_{-j}}$, its own rating r_j, and the ratings of other stores $\boldsymbol{r_{-j}}$. w represents the unit cost of production.

The first-order condition for the optimal price is then derived as:

$$p_j^* - w = -\frac{1}{\alpha(1 - s_j(r_j))}. \tag{7}$$

Here, p_j^* represents the optimal price for store j, w is the unit cost of production, α is the price coefficient in the demand equation, and $s_j(r_j)$ represents the market share of store j given its rating r_j.

During the quality investment stage, sellers optimize their profits by choosing the optimal level of quality investment. The ex post quality of store j is determined by the following equation:

$$\xi_j = e_j + \varepsilon_j. \tag{8}$$

where e_j represents the level of quality investment made by store j, and ε_j represents an error term that affects the realized quality of the product.

Apart from the distribution of investment cost F_c, sellers do not observe the effort nor realized quality of other sellers. The platform sets a rating scheme to identify sellers who achieve the highest quality in the τ-th percentile as H-type. Sellers share a common expectation q_τ regarding the quality threshold necessary to attain this status.

We denote the profits for being ranked as H-type or L-type for firm j as π_H and π_L, respectively:

$$\begin{cases} \pi_H & = D_j(p_j(H), \boldsymbol{p_{-j}}, H, \boldsymbol{r_{-j}}) \cdot (p_j(H) - w) \\ \pi_L & = D_j(p_j(L), \boldsymbol{p_{-j}}, L, \boldsymbol{r_{-j}}) \cdot (p_j(L) - w) \end{cases}.$$

D_j represents the market demand for store j given its price $p_j(r_j)$, prices of other stores $\boldsymbol{p_{-j}}$, and the ratings of other stores $\boldsymbol{r_{-j}}$. w represents the unit cost of production.

The optimization problem during the quality investment stage can be expressed as follows:

$$\max_{e_j} \pi_H Pr(e_j + \epsilon_j > q_\tau) + \pi_L Pr(e_j + \epsilon_j < q_\tau) - c_j e_j^2$$

Here, e_j represents the level of quality investment made by store j, and c_j represents the unit cost of making that investment. We assume a quadratic cost of investment. q_τ is the expected value of the quality threshold.

We assume that ϵ is normally distributed $N(0, \sigma)$, then we can rewrite the decision problem as follows:

$$\pi_j^e := \max_e \pi_L + \Delta\pi\Phi(e - q_\tau) - c_j e_j^2$$

In this equation, $\Delta\pi = \pi_H - \pi_L$ represents the difference in profits between being ranked as H-type or L-type, and $\Phi(\cdot)$ is the cumulative distribution function of ϵ. The variable e represents the level of quality investment made by store j, and q_τ represents the threshold quality level of being ranked as H-type given the threshold of τ.

We can further obtain the first-order condition for the optimization problem as follows:

$$\frac{\partial \pi_j^e}{\partial e} = \Delta\pi\phi\left(e - q_\tau\right) - 2c_j e_j = 0 \tag{9}$$

where $\phi(\cdot)$ represents the probability density function of ϵ.

We assume a symmetric strategy $e(c)$ such that the distribution of realized quality F_q can be defined by the following equation:

$$F(q) = \int_{\epsilon} Pr(e(c) < q - \epsilon | \epsilon) dF_{\epsilon} \tag{10}$$

Here, $F(q)$ represents the distribution of realized quality, $e(c)$ represents the level of quality investment made by sellers, and F_{ϵ} represents the distribution of the error term ϵ in the realized quality equation.

5.3 Equilibrium

In our model, a Bayesian Nash equilibrium is a set of decisions, including pricing choices p_j^*, quality investment decisions e_j^*, and consumers demand decisions D_j^* that satisfy the following conditions simultaneously:

(1) For each seller j, the pricing decision p_j^*, must satisfy the first-order-condition in equation (7), taking into account the firm's own rating r_j, the ratings of all other sellers $\boldsymbol{r_{-j}}$, and the pricing decisions of all other firms $\boldsymbol{p^*_{-j}}$.

(2) Given the profit difference $\Delta\pi(\boldsymbol{p}^*, \boldsymbol{r})$ and the expected quality threshold of attaining H-type status, denoted as q_τ, the optimal level of quality investment e_j^* satisfies the first-order condition in equation (9).

(3) The quality threshold that sellers expect aligns with the threshold resulting from the distribution of realized quality, which is represented as $q_\tau = F_q^{-1}(\tau)$. The equation for F_q is given by (10), and it is dependent on the optimal quality investment $e^*(c)$.

(4) Lastly, consumers accurately anticipate the equilibrium quality distribution F_q and the optimal investment level $e^*(c)$, and they establish expectations for H-type and L-type based on equations (5) and (6).

Our approach for computing the equilibrium closely follows the definition presented above. Given a set of parameters, the steps to compute the equilibrium are as follows:

Outer loop: solve for $\Delta\pi$

1. Make an initial guess of $\Delta\pi$

   Inner loop: given $\Delta\pi$, solve for F_q and e^*

(a) Given $\Delta\pi$ and an initial guess of q_τ, solve for e^*

(b) Given e^*, obtain $F_q(e)$ and solve for q'_τ

(c) check if $q_\tau = q'_\tau$. If not, go back to step (a)

2. Given F_q, solve for $E[\xi|H]$ and $E[\xi|L]$
3. Given $E[\xi|H]$ and $E[\xi|L]$, solve for p^*_H and p^*_L
4. Calculate $\Delta\pi'$, check if $\Delta\pi = \Delta\pi'$. If not, go back to step 1.

6 Conclusion

In conclusion, our collaboration with one of the world's largest business-to-consumer online retail platforms has enabled us to utilize the extensive data captured on seller operations to design and generate reliable platform-generated ratings. These ratings serve as a valuable tool for customers in making informed purchase decisions while also promoting greater transparency and trust in the online retail marketplace.

Online Appendices: Not For Publication

A Calculation of Platform-generated Rating

- **Score in Each Dimension**: The platform compares the seller performance in each dimension with the performance of other sellers in the same category and assign 5, 4, 3, 2, 0 subscores to each dimension. PGR measures the relative performance of the sellers. For instance, suppose the seller respond rate is top 15% in the category, seller obtain a subscore of five in this dimension.

- **PGR Score**: Calculate a weighted average over all subscores and obtain a PGR score. Usually the platform assign equal weight to all five subscores but there are variation across categories when the platform believes some dimensions are more important than others.

- **PGR Star**: Finally The platform Match PGR score to PGR star. Table A.1 summarizes the relationship. Sellers obtain a higher star levels when they have higher star scores. It is worth noting that Star level 4.5 and 5 have the same minimum score but sellers have to satisfy additional service requirement such as instant refund to be eligible to 5 star.[9]

Table A.1: Relationship between New Light House Star Score and Star Level

Star Level	Star Score	5 Star Requirement
3	≤ 2.4	
3.5	≥ 2.5	
4	≥ 3.6	
4.5	≥ 4.1	
5	≥ 4.1	Yes

[9] In general, the score is a minimum requirement for the star level. There are additional requirements such as GMV level and no punishment from the platform.

B PGR Transition Matrix

Next, we summarize the score distribution and transition probability in the data. Table B.1 reports the transition matrix.

Table B.1: Star Level Transition Matrix

	Star (current)					
Star (lag)	3	3.5	4	4.5	5	Total
3	23744	31136	2263	85	13	57241
3.5	35018	1147479	395181	13862	3219	1594759
4	1531	402702	1625722	159962	49234	2239151
4.5	56	12515	165171	317968	9240	504950
5	5	2625	49317	6560	147136	205643
Total	60354	1596457	2237654	498437	208842	4601744

C Demand Estimation: Saturated

Table C.1: Transition Matrix: Saturated (# Visitors)

	Star(current)			
Star (lag)	3	3.5	4	4.5 or 5
3	0	0.00807***	0.00670	0.0282
		(0.00308)	(0.01)	(0.0413)
3.5	-0.0155***	0	0.00293***	-0.00634***
	(0.00190)		(0.000581)	-0.0024
4	-0.0309***	-0.00563***	0	0.00263***
	(0.0103)	(0.000484)		(0.00059)
4.5or 5	-0.0871**	-0.0118***	-0.00431***	0
	(0.0434)	(0.00247)	(0.000595)	

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table C.2: Transition Matrix: Saturated (# Purchase)

	Star (current)			
Star (lag)	3	3.5	4	4.5or 5
3	0	0.0176***	-0.00192	-0.00256
		(0.00631)	(0.0186)	(0.0804)
3.5	-0.0407***	0	0.0135***	-0.00180
	(0.00391)		(0.00115)	(0.00531)
4	-0.0672***	-0.0247***	0	0.0294***
	(0.0188)	(0.00109)		(0.00140)
4.5	-0.0247***	-0.0608***	-0.0403***	0
	(0.0765)	(0.00567)	(0.00164)	

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table C.3: Transition Matrix: Saturated (GMV)

	Star (current)			
Star(lag)	3	3.5	4	4.5 or 5
3	0	0.0171	-0.0106	-0.133
		(0.0194)	(0.0514)	(0.243)
3.5	-0.0339***	0	0.0229***	0.0201
	(0.0124)		(0.00300)	(0.0158)
4	-0.0539	-0.0254***	0	0.0369***
	(0.0697)	(0.00288)		(0.00321)
4.5 or 5	-0.504	-0.0223	-0.0429***	0
	(0.312)	(0.0187)	(0.00367)	

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table C.4: Demand Estimation: Saturated (CVR)

	Star (current)			
Star (lag)	3	3.5	4	4.5 or 5
3	0	0.00122***	-0.000604	0.00489
		(0.000397)	(0.00239)	(0.00462)
3.5	-0.00153***	0	0.00139***	-0.000432
	(0.000307)		(0.000494)	(0.000638)
4	-0.00299**	-0.00198***	0	0.000200
	(0.00132)	(0.000464)		(0.000937)
4.5	-0.0134**	-0.00363***	0.00252***	0
	(0.00560)	(0.000308)	(0.000111)	

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

D Demand Estimation: Robustness

Table D.1: Demand Estimation: Regression Discontinuity (Three Day Difference)

	(1) Δln(# of Visitors)	(2) Δln(# of Buyers)	(3) Δln(GMV)	(4) ΔCVR
ΔStar	0.00267***	0.00890***	0.0155***	0.000825***
	(0.000402)	(0.000920)	(0.00249)	(0.000173)
Constant	-0.0151 *	0.0783***	0.110***	0.00256***
	(0.000603)	(0.00140)	(0.00314)	(7.14e -05)
Observations	4,601,744	4,601,744	4,601,744	4,600,836
R-squared	0.055	0.264	0.083	0.002

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table D.2: Demand Estimation: Regression Discontinuity (Five Day Difference)

	(1) Δln(# of Visitors)	(2) Δln(# of Buyers)	(3) Δln(GMV)	(4) ΔCVR
ΔStar	0.0113***	0.0152***	0.0226***	0.000136**
	(0.000547)	(0.000847)	(0.00141)	(5.78e-05)
Constant	-0.00865***	0.00643***	0.00424***	7.26e-05
	(0.000811)	(0.00116)	(0.00168)	(6.26e-05)
Observations	4,601,744	4,601,744	4,601,744	4,601,662
R-squared	0.145	0.280	0.180	0.002

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

E Demand Estimation: UGR

We estimate the demand-side responses from UGR (same specification as the PGR demand estimates). We find that UGR is associated with a higher number of visitors but lower conversion rates and the overall impact on GMV is not significant. This may come from the fact that UGR varies little from day to day. The most variation comes from a 0.01 change in UGR.

Table E.1: Demand Estimation with UGR Description: (One Day Difference)

	(1) Δln(# of Visitors)	(2) Δln(# of Buyers)	(3) Δln(GMV)	(4) ΔCVR
ΔUGR Description	0.2111***	-0.1393	-0.2922	-0.0663***
	(0.037)	(0.1003)	(0.2778)	(0.0225)
Observations	4,601,671	4,601,671	4,601,671	4,600,764
R-squared	0.055	0.264	0.828	0.124

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table E.2: Demand Estimation with UGR Logistics (One Day Difference)

	(1) Δln(# of Visitors)	(2) Δln(# of Buyers)	(3) Δln(GMV)	(4) ΔCVR
ΔUGR Logistics	0.1325***	-0.1574	-0.347	-0.065**
	(0.0383)	(0.0979)	(0.3061)	(5.78e-05)
Observations	4,601,671	4,601,671	4,601,671	4,600,764
R-squared	0.0553	0.264	0.080	0.124

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.

Table E.3: Demand Estimation with UGR Service (One Day Difference)

	(1)	(2)	(3)	(4)
	Δln(# of Visitors)	Δln(# of Buyers)	Δln(GMV)	ΔCVR
ΔUGR Service	0.1450***	-0.2824***	-0.4884	-0.0678***
	(0.0369)	(0.0956)	(0.0828)	(0.0235)
Observations	4,601,671	4,601,671	4,601,671	4,600,764
R-squared	0.0553	0.264	0.083	0.124

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Control variables: period fixed effect.