# Heterogeneous Effects of Online Reputation for Local and National Retailers*

Peter Newberry
*Department of Economics*
*The Pennsylvania State University*

Xiaolu Zhou
*School of Economics and Wang Yanan Institution for Studies in Economics*
*Xiamen University*

November 19, 2018

**Abstract**

We study the heterogeneous effect of online reputation for sellers that differ in their national presence and examine how this heterogeneity affects the distribution of sales on a large Chinese platform. We estimate a demand model that incorporates a structural learning process and allow for the process to vary across sellers who are differentiated by their national presence. The estimates suggest that the impact of reputation is larger for local sellers. Using these estimates, we find that removing the reputation system would result in large shift of demand from local sellers to national sellers. This shift is due to the fact that high quality local sellers can no longer differentiate themselves from low quality sellers in the marketplace.

**Keywords:** Reputation, E-commerce, Word-of-Mouth, Online reviews, Online Competition
**JEL Codes:** D83, L15, L81

*This research is supported by the National Natural Science Foundation of China (grant 71803162). We'd like to thank Paul Grieco, Charles Murry, Mark Roberts, Chris Parker, and Xiang Hui for their helpful comments and suggestions. All correspondence may be addressed to the authors via e-mail at `pwnewberry@psu.edu` or `xiaoluzhou@xmu.edu.cn`

# 1 Introduction

To lessen the impact of information asymmetries that are due to the impersonal nature of the internet, most online markets have peer review systems that provide consumers a signal of a seller and/or product's quality (i.e., reputation).[1] Survey and empirical evidence suggest that these systems are an important part of a consumer's decision-making process.[2] At the same time, there exists heterogeneity in online sellers in terms of their experience, their size, their prevalence in offline markets, etc. In this paper, we study the variation in the impact of reputation across heterogenous sellers in an online marketplace.

Many of the sellers on Alibaba's Tmall (our data source) are small sellers who only have a local offline outlet, while other sellers are large manufacturers or well-known retailers who have an extensive offline presence. Absent any measures of reputation, consumers may concentrate their demand on the offline firms with whom they have interactions outside of Tmall, making it difficult for lesser-known firms to be able to compete in the marketplace. However, it may also be the case that consumers view the online version of the national retailer as different from the offline version, implying that the role of online reputation might not differ across this dimension. In this paper, we ask: to what extent does an online reputation system determine the distribution of demand across sellers who differ in their national presence?

To answer this question, we quantify the impact of the rating system on Tmall, a branch of China's largest e-commerce company, Alibaba. Tmall is a business-to-consumer (B2C) platform which features thousands of professional sellers offering a wide variety of products. While there have been numerous previous papers studying Taobao, Alibaba's consumer-to-consumer marketplace, this is the first paper to use data from Tmall to the best of our knowledge. This platform is a leader in China's online B2C market, as it had 54 percent market share and total transactions reached $39 billion in Q3 2015.[3] Tmall features a rating system that is similar to that of Amazon.com,

[1] For example, Amazon.com displays the distribution of seller and product ratings (1 through 5) given by previous shoppers, while ebay.com has a feedback system in which users indicate whether their experience was positive or negative.

[2] For survey evidence see `http://marketingland.com/survey-customers-more-frustrated-by-how-long-it-takes-to-resolve-a-customer-service-issue-than-the-resolution-38756`. Both Chevalier and Mayzlin (2006) and Dellarocas (2003) provide empirical evidence.

[3] Information from `http://www.chinainternetwatch.com/15957/chinas-b2c-sales-q3-2015/`.

where each customer purchasing a product from a given seller rates the quality of the transaction on a scale of 1 to 5. Customers who arrive thereafter can observe both the average rating score and the total number of ratings for a given seller.[4]

Our data include monthly prices and quantities for tablets sold on Tmall between September, 2014 and April, 2015. We observe both product and seller characteristics, where the latter include the rating score (i.e., average rating) and the complete distribution of ratings. Additionally, we observe a classification of sellers, as defined by Tmall, which is based on a seller's national presence. The classification allows us to separate retailers into two groups which we call 'types': those who are well-known nationally (hereafter 'national sellers') and those who are local or online only retailers (hereafter 'local sellers'). Our aim is to estimate the impact of the rating system in determining the concentration of sales across these two types.

We do this in three steps. First, we show descriptive evidence that seller ratings have a larger impact for local sellers. When accounting for product characteristics and utilizing the covariation between ratings and sales within a seller, we find evidence that ratings positively affect demand for local sellers but not national sellers.

Next, we estimate a discrete choice model of demand where the consumer chooses to buy a tablet from one of the sellers on Tmall, but is uncertain about the quality of the sellers prior to making the purchase. This uncertainty is a key dimension of differentiation in online markets, as many sellers offer similar products. Therefore, we enrich the demand model by assuming that the consumer infers the expected quality of each seller utilizing the average rating and the number of ratings and a Bayesian learning process. We allow the updating process (i.e., the learning parameters) to be functions of the seller's type, along with other seller characteristics. The differences in the learning parameters across seller type determines the extent to which there is heterogeneity in the effect of ratings.

The results show that consumers have a higher prior belief about national sellers, implying that consumers believe that these sellers are of higher quality, *ex ante*. Additionally, we find that the weight placed on the rating score is significantly higher for local sellers, meaning the rating score

[4]See Figures 1 and 2 for screen shots of a search page and a seller's home page.

has a larger effect on demand for these sellers. This suggests that consumers rely more heavily on their prior belief about quality rather than the rating score for national sellers, whereas they use the rating score to infer quality for local sellers.

Finally, using the estimates of the model, we quantify the effect of the reputation system on demand across seller types by removing ratings and forcing consumers to use only their prior beliefs about seller quality to make their purchase decision. When removing ratings and keeping prices and all other aspects of the market fixed, we find that the total market share of national sellers increases from 20% to 39%, while the total market share of local sellers decreases from 66% to 31%.[5] This implies that the lack of ratings leads to an increase of the outside share, but the primary effect is to redistribute sales between the two different types, as buyers now focus their sales on national sellers despite the fact that they might find local sellers who are of higher quality. We see the largest effects for the biggest sellers, as the total market share of the top 20% of national sellers increases from 17% to 35% and the total market share the for top 20% of local sellers drops from 61% to 27%. This implies that the reputation system on Tmall allows the 'best' local sellers to compete with, and even outsell, their national counterparts.

To take a closer look at the substitution between local and national sellers, we calculate the effect of removing ratings across different quality groups for each type, where a quality group is defined by the sellers' rating score at the end of the sample. We find that removing the rating score increases market share for national sellers at all levels of ratings, even those with a perfect rating score. On the other hand, the lowly-rated local sellers are better off and the highly-rated local sellers are hurt by the removal of ratings, suggesting that the ratings allow the high quality local sellers to compete with the national sellers, and at the same time, reveal information about the low quality sellers. This highlights the importance in accounting for the heterogeneity in the effect of ratings, as homogenous learning parameters would result in highly-rated sellers being hurt by the removal of ratings regardless of seller-type.

Overall, the results suggest that the online reputation system on Tmall plays an important role in determining the nature of competition on the platform. The system serves to alleviate

[5] In a robustness check, we have computed all exercises allowing sellers to have a price response via a reduced form pricing function and the results change very little. See Section 5 for details.

information asymmetries between buyers and relatively unknown local sellers on the platform, which allows the high quality local sellers to compete with the more well-known national sellers. As retailers such as Target and Wal-Mart continue to focus efforts on increasing their online presence and Amazon continues to dominate the online market, it becomes more important to understand how and why competition occurs in these types of markets. Our analysis highlights the role of online reputation systems in allowing 'mom and pop' shops to compete with these large retailers.

This paper is related to a couple of different strands of literature. First, there are a number of papers which estimate the effect of online reputation on demand. Chevalier and Mayzlin (2006) estimate the effect of reviews at the product level using data from book sales on Amazon, while Dellarocas (2003), Cabral and Hortacsu (2010), Nosko and Tadelis (2015) and Saeedi and Sundaresan (2016) study the impact of seller reputation on eBay. Zhu and Zhang (2006) and Su et al. (2016) do the same but focus on Taobao. Zhu and Zhang (2010) estimate the impact of online reviews in the video game industry and find that the effect of reviews varies across games that differ in their characteristics. Elfenbein et al. (2015), Hollenbeck (2016), Fang (2018), and Luca (2011) have a similar goal to this paper, but instead of looking at heterogeneity in terms of an online seller's national presence, they examine the effect of reviews across sellers who differ by their experience or chain affiliation. In other words, we differ in that we examine the role of a retailer's brand awareness in determining the impact of reputation in an online marketplace, where much of the brand awareness is coming from interactions that occur outside of the marketplace. Additionally, we connect ratings and reputation to demand through a structural learning model, allowing us to estimate a rich set of learning parameters. Similarly, Zhao et al. (2013) estimate a structural learning model where expected quality of a book's genre is based on user ratings, but do so for a single seller.

Additionally, our paper is related to the large amount of industrial organization literature studying demand for differentiated products that follow Berry (1994) and Berry et al. (1995). That is, we construct a discrete choice model under incomplete information, where the incomplete information is in terms of seller quality. We allow consumers to learn about seller quality through a reputation signal (i.e., ratings) rather than through costly search. Specifically, we add structure

to the standard unobserved product quality term, allowing it to be formed based on a Bayesian learning process where consumers update their belief using the online reputation system. Grennan and Town (2015) and Chernew et al. (2008) model learning in a similar manner, but do so in the health care industry.

The rest of the paper is organized as follows. Section 2 discusses the relevant institutional details and introduces the data, while Section 3 specifies the demand model. The estimation of the model is in Section 4, the results are presented in Section 5, and Section 6 concludes.

## 2 Data

Alibaba Group Holding Limited, launched in 1999, is the leading online commerce provider in China. While the company offers a broad spectrum of e-commerce services, the two primary platforms are taobao.com and tmall.com. Taobao is similar to eBay, as it is an open platform connecting individual sellers and buyers, also known as a consumer-to-consumer (C2C) market. On the other hand, Tmall specializes in providing a marketplace for professional sellers and enterprises, meaning it resembles Amazon Marketplace, a business-to-consumer (B2C) market. These two platforms sell a wide range of products, including apparel, shoes, books, electronics, smartphones and televisions. Their sales account for nearly 80% of China's total online shopping market, and their gross merchandise volume is the largest in the world.[6] Total sales reached 2,950 billion RMB (around $466 billion) in 2015, with $285.78 billion on Taobao and $180.25 billion on Tmall.[7] For comparison purposes, eBay's sales from 2015 were $81.7 billion.[8]

Much like other e-commerce platforms, Tmall has a system that allows consumers to rate their experience with a given seller/product. After each transaction, the buyer is required to post a rating score for three aspects of the experience: (1) the seller's descriptions of goods, (2) the customer service, and (3) the shipping service. Scores range from 1 to 5, with 1 being the lowest score in terms of satisfaction. The average rating score (rating score, hereafter), which is the primary information

[6] Statistics can be found at `http://www.alibabagroup.com/en/ir/glance`.

[7] Information is from Alibaba's financial report, downloaded from `http://www.alibabagroup.com/en/ir/financial`.

[8] Information is from eBay's annual financial report downloaded from `https://investors.ebayinc.com/annuals.cfm`.

that is displayed to consumers, is calculated based on the posted ratings for all transactions that occurred in the previous six months.

In order to observe the rating score along and the total number of ratings in the previous six months, an individual must either display the search results 'by seller' by clicking a link on the default search page, navigate to a page that displays all sellers selling a particular product, or go to an individual seller's page. In the former two scenarios, the consumer would observe the information for many sellers at once, while in the latter, she would only observe the information for one seller. An example of a search result sorted by seller is shown in Figure 1. The top listed seller, Kindle Official Store, has rating scores of 4.9, 4.8, and 4.8 for their description, customer service, and shipping services, respectively, and the first Kindle listed has a total of 8,253 ratings in the previous six months.

If a customer clicks to a specific seller's page, she can observe more detailed records about a seller's reputation. Specifically, the distribution of rating scores given to the seller and the total number of ratings for that store in previous six months are displayed. Figure 2 shows that the Kindle Official Store's customer service has been rated by 22,887 consumers and that 94.11% (21,539) of the ratings are the highest and 0.52% of the ratings are the lowest.

Another important feature of Tmall is that it employs a classification system of sellers. Sellers are separated into three different groups that are based on the authorization contract with the product's manufacturer and the scale of the business. Group-one sellers are the legal representatives of a registered trademark, which could be a product brand such as Apple (iPad), Amazon (Kindle), or Samsung (Galaxy Pad). Additionally, large retailers with a registered store brand (e.g., Best Buy) could be invited to be a group-one seller.[9] Group-two sellers are enterprises which are authorized by the brand owners to sell their products, meaning they typically sell only one brand. Finally, for the most part, group-three sellers are authorized retailers who sell multiple brands that they purchase directly from suppliers.

Table 1 displays information on the number of offline stores, the number of employees, and the

[9]We use retailers popular in the United States as examples to fix ideas. These retailers do not necessarily operate on Tmall.

existence of a website separate from Tmall for the top five sellers of each group.[10] Note that some of this information was not available for every group. The primary difference between group-one sellers and the other two groups is in the store's brand recognition through interactions occurring outside of the Tmall platform. Group-one sellers generally have a large offline presence and have their own website, meaning they are likely well-known to customers throughout China. Thus, we call them 'national' sellers. These measures imply that group-two and three sellers are similar in size and usually lack a website. Because of this, we join group-two and three into one group, which we call 'local' sellers. These sellers generally have either a small local presence or are retailers who sell only online.

### Seller Descriptive Statistics

The data include all sales of tablets which occurred between September, 2014 to April, 2015. We collect the data by scraping the Tmall website on the 11th or 12th of each month, depending on how many days occur in that month. For each seller who sells tablets on Tmall, we observe the quantity sold in the previous month and the price at the time of the scrape for all the products which it sells. We formulate the price by taking the average of the two prices: the one observed last month and the one observed this month. If the product was not sold by the seller last month, we use only this month's price. We also observe the value of the three different rating scores and the number of ratings in the previous six months at the time of the scrape for each seller. Finally, we observe the following product characteristics for each product: brand, operating system, screen size, memory, storage, cellular internet capability, and the number of years since the product's introduction (i.e., product age).

The market for tablets on Tmall is described in Table 2. The first two columns display statistics for all the sellers offering tablets, while the last two columns show only sellers who are identified as electronics retailers. The non-electronics retailers are general merchandise stores selling goods in many different product categories. Moving forward, we focus our analysis on the electronics

[10] This information was gathered by searching for each seller's official website where, often times, the number of offline retail outlets are displayed. When the seller did not have an official website, we search for the store in a business directory that lists the number of employees.

retailers and assume that buying from a general merchandiser is the 'outside good'. This allows us to limit the amount of heterogeneity across sellers to a certain extent.

For the eight month sample period, there were 670 different sellers, with a majority of those (562) being electronics retailers. The average number of electronics sellers on the platform per month is about 391. There were 491 total local sellers and 71 national sellers within the electronics category, making up 87% and 13% of total sellers. The table also shows that there are many different brands and models sold on Tmall. Finally, Tmall is a very large seller of tablets with nearly 83 thousand units sold and $20 million in revenue per month for electronics retailers alone.

Next, we examine the heterogeneity across the different seller types in Tables 3 and 4. Table 3 presents seller-level descriptive statistics separated by seller type. The average quantity sold and revenue over the eight month time period for national sellers are 2,125 and $444,840, but these distributions are heavily skewed with the medians being 171 and $36,768. The local sellers are smaller, as the median local seller sells less than half as many tablets as the median national seller. Additionally, local sellers have prices that are about $20 higher than national sellers. However, the sellers do not significantly differ in their scope, as the number of brands and the number of models are similar across the different types of sellers. Finally, the different types of sellers do not significantly differ in their tenure, where tenure is measured by how long the seller has been on Tmall.

Table 4 presents the average characteristics for the tablets sold by each type in the first two columns and across all seller-types in the last column. While some brands familiar in the United States are also popular on Tmall (e.g., Apple is #4 and Samsung is #9), the top three brands are Teclast, Onda, and Cube. Android and Windows operating systems appear on nearly 61% and 33% of tablets, respectively, while iOS is on about 10% of the tablets sold. Note that there are some tablets which have multiple operating systems available, resulting in the total across the three major ones being over 100%. Also, there are a few lesser-known operating systems which appear in the data, but account for a very small percentage of the market. About 26% of the tablets have internet access through cellular networks, the average screen size is 8.5 inches, the average memory is just under 2 GB, and the average storage is around 28 GB. Finally, the average tablet is sold

about 7 months after its initial release date.

Now focusing on the first two columns, it is apparent that sellers do not differ much in the characteristics of the tablets they offer, but do seem to differ somewhat in the brands that they sell. Despite this, the descriptive statistics suggest that national and local sellers are similar along many of the dimensions we examine. While we account for these dimensions in our analysis, this fact provides reassurance that any differences in the effect of ratings across seller type are not primarily driven by heterogeneity in the types of goods offered across sellers.

Finally, we describe the distributions of ratings and market share across seller type in Tables 5 and Table 6. The first three rows of Table 5 show the mean, median, and standard deviation for the distributions of rating scores for each seller type. Note that the rating score we present is the average of the three different rating scores (i.e., description, service, and shipping). The correlation of the three different scores is as low as 0.92 and as high as 0.97, so presenting them separately would result in a similar pattern.

The mean rating scores are 4.72 and 4.70 and the standard deviations of the rating scores are 0.26 and 0.32 for local and national sellers, respectively. So, while rating scores are skewed towards the top, there is some variation in scores between 4 and 5 for both types of sellers. While these moments of the distribution of ratings are similar across type, it does not necessarily imply that the distributions are the same. To examine this closer, we perform a Kolmogorov-Smirnov test of equal distributions of ratings across seller type and present the results in Table 6. The results show that the null hypothesis of equal distributions can be rejected, suggesting that the ratings vary across seller type. Further, the second row of Table 6 suggests that the distribution of ratings for national sellers stochastically dominates that of local sellers, meaning that national sellers have statistically higher rating scores.

The middle panel of Table 5 presents descriptive statistics on the number of ratings across types. It shows that the average and median number of ratings for the national sellers are significantly higher than for the local sellers, which comes from the fact that national sellers have a higher sales volume. However, there is a substantial amount of variation in the number of ratings for both types of sellers, meaning that there are sellers of each type that have few ratings. This allows for the

identification of the effect of ratings, as the rating score can vary significantly for these sellers.

The last panel of Table 5 presents descriptive statistics on the market share across type. The first row displays the total market share, averaged across the eight months of our data, which demonstrates that local sellers dominate in this dimension. However, this is due to the fact that there are so many more local sellers, as can be seen by the average market shares presented in the second row.

The fact that national sellers have both higher market share and higher ratings provides preliminary evidence that these sellers are, in general, higher quality sellers. However, seller quality can vary within seller type, meaning that some national sellers may be lower quality than the average local seller and some local sellers may be higher quality than the average national seller. The rating system on Tmall provides a way for consumers to decipher between these quality levels through the experiences of others. But how these experiences affect the beliefs of consumers, and ultimately demand, depends on how informative these experiences are in shaping consumers' beliefs. The goal of the model presented in Section 3 is to estimate how consumers form their beliefs about quality through ratings and to examine how this belief formation process may differ across seller types.

### Preliminary Evidence

Before presenting the formal model, we provide descriptive evidence that the effect of ratings is stronger for local sellers than for national sellers. To do this we run an OLS regression of log monthly sales on product characteristics, a seller fixed effect, and a function of the rating score and the number of ratings. Importantly, this function varies by seller in terms of whether or not they are a national or local seller. The specification is given by:

$$log(Q_{jst}) = \beta X_{jst} + \xi_s + f_s(r_{st}, n_{st}) + \epsilon_{jst} \tag{1}$$

where $Q_{jst}$ is the quantity sold of product $j$ by seller $s$ during month $t$. The vector $X_{jst}$ contains a constant, the log of the price, and product characteristics including screen size, storage, memory, the length of time the tablet model has been on the market, a dummy variable indicating whether or not the tablet has access to wireless internet networks, a set of operating system dummy variables,

a set of brand dummy variables for the top 15 brands and a month fixed effect. The seller fixed effect is given by $\xi_s$.

For this exercise, we assume a linear function of the average rating score, $r_{st}$, and the number of ratings, $n_{st}$:

$$f_s(r_{st}, n_{st}) = \alpha_{1s} r_{st} + \alpha_{2s} n_{st} + \alpha_{3s} n_{st} r_{st}$$

where we allow the $\alpha$s to vary by whether or not the seller is a national or local seller. The rating score and the number of ratings are equal to their values measured at the beginning of month $t$. Recall, that these values are calculated by Tmall based on the ratings for the seller in the past six months. Also note that the average rating score is the average across the three different rating scores. This function is a reduced form representation of a learning model, as it takes into account both a 'base effect' of the rating score and an interaction effect, allowing for the rating score to become more informative as the seller receives more ratings.

In Figure 3, we display the estimate of the effect of the rating score over different values of $n_{st}$ for both types of sellers and the 95% confidence interval for these estimates. The y-axis in this figure represents the percentage change in demand for an increase in 0.1 in the rating score. The figure shows that the rating score has a significant positive effect on sales for local sellers at all levels of $n_{st}$, whereas the effect for national sellers is positive, but not significantly different from 0. Additionally, the point estimate of the effect of the rating score increases for local sellers as $n_{st}$ increases, but stays relatively steady for national sellers.

Overall, these results suggest that ratings are more informative, in terms of inferring the quality, for local sellers than they are for national sellers. One worry may be that this is due to the fact that local sellers sell different products than national sellers and that the effect of ratings is actually heterogeneous across products, rather than sellers. For example, it may be the case that ratings are informative for relatively unknown brands and that local sellers are more likely to sell these brands. The descriptive statistics discussed in Section 2 demonstrate that the latter is likely not the case, but we check for heterogeneity in the effect of rating across brands in Figure 4.

The figure displays results of regressions similar to that of the specification in Equation 1, but we allow the effect of ratings to vary by both seller type and by whether or not the brand is well-

known. We group products into well-known and unknown brands based on the brand's sales on Tmall during our sample period. Specifically, the well-known brands are the top five selling brands for the eight month period in our data. In an alternative specification, we define a well-known brand as one with a top five level of 'brand recognition', as defined by the authors, and the results are very similar.

The results indicate that the effect of reviews is consistent across well-known and unknown brands for each type. For national sellers, the effect of reviews is insignificant for both types of products, and for local sellers it is significant and positive for both types of products. This reduced form evidence suggests that the primary heterogeneity in the effect of reviews is across seller types rather than across product characteristics (i.e., brand).

Previous literature has found that the effect of online reputation varies across sellers of different experience and chain affiliation. This result suggest that the effect of reputation also varies across sellers who differ in terms of their national presence via offline sales and brand recognition. From this, one could infer that local sellers rely relatively more on the online reputation system in order to compete in the online marketplace. In order to speak to this, we specify and estimate a structural model of demand in which we allow consumers to learn about seller quality through the reputation system on Tmall. We then quantify the impact of the reputation system on the distribution of demand across heterogenous sellers. The model is formulated in the following section.

# 3 Model

We specify a model of demand that incorporates consumer learning from seller ratings. Before making her purchase decision, we assume that the consumer observes the three rating scores and the total number of ratings in the previous six months for each seller. While the rating scores don't appear on the default search page, they do appear when an individual sorts by seller within the search results. Additionally, if a consumer is looking for a particular tablet, she will navigate to a page displaying all the sellers selling that tablet, along with the rating score information. Finally, once a consumer navigates to a seller's page before purchasing the tablet, the information on the entire distribution of ratings is visible.

While the total number of ratings is only visible once the consumer clicks to the seller's home page, the number of ratings for a seller/tablet pair are visible on the search page sorted by seller. To the extent that we see the learning process as an approximation of the observed behavior, it is reasonable to assume that consumers can infer the accuracy of the average rating from the total number of ratings for the product(s) which appear on the search page. However, taken literally, the assumption implies that consumers are made aware of the total number of ratings by clicking to each of the seller's home pages.

While there are three rating scores displayed for each seller, we assume that the consumer uses the average of the three scores to form her belief about seller quality. As mentioned in Section 2, the three scores are highly correlated, meaning forming a belief based on one score is approximately equivalent to forming a belief based on all three. In addition, we performed the reduced form analysis from Section 2 separately for the three different scores, and we find similar results across these specifications.

Finally, we assume that consumers know the ratings information for all sellers. One may worry that any variation in the effect of ratings we estimate may be due to the fact that potential customers do not observe the information for all sellers. However, as long as the observable information doesn't systematically vary across seller-type, then we will still be able to estimate the *difference* in the effect of rating across this dimension. For example, it may be the case that consumers look at only the sellers on the first few search pages and decide which seller to buy from among this subset of sellers. But as long as there is an exogenous mix of national and local sellers on these pages, then the estimated *difference* in the effect of ratings across seller type will not be a function of this misspecification error. However, in this case, the *level* of the effect of ratings for both seller-types would be underestimated, as we assume the effect of ratings is the same for sellers who are highly visible and those who are not. This is because we would be attributing the lack of a relationship between demand and ratings for less visible sellers to a smaller effect of reputation, when it is actually due to the fact that consumers are not considering those sellers.

We address this issue in two ways. First, we show that sellers are not systematically sorted based on seller-type in the default search. To do this, we document the percentage of national

sellers that appear on the first few search pages in a default search undertook in January of 2018. Note that there are 60 listings per page in the default search. We found that the percentage of national sellers is approximately 35-45% through page 5 (top 300 listings), and this percentage remains relatively constant from page 1 to 5. This suggests that there are a significant number of the different seller-types across all the search pages. Additionally, using these same data, we regress the rank of a seller/product in the default search on its price, its quantity of sales in the past month, a dummy variable indicating the seller's type, and interactions between the seller-type dummy and the other covariates. Results show that rankings are based on price and quantity but not on seller-type or the interactions. This indicates that the default search rank is not significantly biased towards any seller-type.

Second, we provide evidence that the effect of ratings is robust to different assumptions about the extent to which consumers have limited choice sets. To do this, we perform the reduced form analysis on the different subsets of our data, where the subsets were defined based on the ranking of products in the default search results. Unfortunately, in our data, we do not observe where the listing appears in the default search order, so we use an 'educated guess' of how the listings would have been sorted. The guess is the predicted ranking based on the estimated parameters of a rank-order regression similar to the one discussed in the previous paragraph, but without the seller-type effects, as these were found to be insignificant. The analysis suggests that the effect of ratings differs across seller-type when limiting to listings that appear in the first five pages, and that the difference in this effect remains as we include more pages. Further, the estimated *level* of the effect for local sellers and the significance of the effect for national sellers does not change when including more pages. Put together, this provides reassurance that both the differences and the levels of the effect of ratings across sellers isn't coming from any systematic ordering of products in the default search.[11]

[11]To the extent that there are other unobserved reasons why local sellers may be considered more often than national sellers, and those reasons are correlated with ratings, then our model will produce biased estimates of the effect of ratings in both differences and levels. However, we do not have a reason to believe that this is the case.

### Demand Model

We assume that $M_t$ consumers arrive in month $t$ and choose to purchase a tablet from one of the electronics retailers on Tmall or to purchase from a seller who is not an electronics retailer (option 0). Consumer $i$'s utility of purchasing product $j$ from store $s$ in month $t$ is:

$$U_{ijst} = \beta X_{jst} + \xi_s + \nu_{jst} + \epsilon_{ijst} \tag{2}$$

where $X_{jst}$ is the same vector as the reduced form analysis, $\xi_s$ is the true quality of the seller, $\nu_{jst}$ is a product/seller-level demand shock, and $\epsilon_{ijst}$ is an idiosyncratic taste shock assumed to be iid extreme value. We assume that $\xi_s$ is fixed, meaning stores can not adjust their true quality over our sample period. As it is only eight months, it is reasonable to assume that it is difficult to adjust quality in the short run. Finally, we normalize the mean utility of the outside option to 0:

$$U_{i0t} = \epsilon_{i0t} \tag{3}$$

Consumers face uncertainty about the quality of a given seller prior to purchase, implying that they make purchase decisions based on expected utility:

$$E[U_{ijst}] = \beta X_{jst} + \underbrace{E[\xi_s | r_{st}, n_{st}]}_{\zeta_{st}} + \nu_{jst} + \epsilon_{ijst} = \beta X_{jst} + \zeta_{st} + \nu_{jst} + \epsilon_{ijst} \tag{4}$$

where $\zeta_{st}$ is the expected store quality given the average rating score, $r_{st}$, the number of ratings in the previous six months, $n_{st}$, measured at the beginning of month $t$. We assume that the expected store quality is formed via a Bayesian learning process which we describe in the following section.

Given the assumption of the errors, the probability that consumer $i$ purchases product $j$ from seller $s$ is:

$$P_{ijst} = \frac{exp(\beta X_{jst} + \zeta_{st} + \nu_{jst})}{1 + \sum_{j,s} exp(\beta X_{jst} + \zeta_{st} + \nu_{jst})}$$

We do not assume any heterogeneity in preferences, meaning this maps directly into the market

share of each product:

$$s_{jst} = \frac{exp(\beta X_{jst} + \zeta_{st} + \nu_{jst})}{1 + \sum_{j,s} exp(\beta X_{jst} + \zeta_{st} + \nu_{jst})} \tag{5}$$

### Bayesian Learning Model

In order to specify the learning model, we first discuss how ratings are generated. After purchasing a tablet, each customer is required to rate the transaction. If a rating is not posted, a 5 is automatically given. We account for this in a robustness check and neither the estimates nor the results of the simulations vary to a great degree.[12] We assume that a customer posts a rating based on her shopping 'experience', which is generated from a distribution centered around the true seller quality. It is important to emphasize the fact that the reviewer does not intend to reveal her *belief* about the quality of the seller, but rather she wishes to share the quality of her *individual shopping experience*.

The independence of signals implies that, while consumers can learn about the seller's true quality through ratings, each rating is not a function of the history of previous ratings. This simplifies the estimation procedure substantially, as it implies that consumers (and the econometrician) do not have to integrate over all the possible combinations of ratings that would have led to the observed average as in Newberry (2016). Instead, the consumer sees the rating score as an average of independent draws from the signal distribution and calculates the updated belief based on these draws.

Not only does this assumption make the model more tractable, but we also believe it is realistic in online markets which feature review systems. That is, suppose that a customer has a very bad experience with a seller. It is reasonable to believe that she will give a very bad rating irrespective of the previous rating score. To further justify this assumption, a reviewer cannot observe the ratings of her peers or the average rating while she is submitting her own, making it less likely that the number she gives is state-dependent.

More specifically, we assume that shopping experiences arrive as a random signal measured in

[12]Specifically, we assume that consumers discount (i.e., ignore) approximately 47.9% of the perfect ratings. This number comes from Dellarocas (2003), which states that 52.1% of consumers take time to leave feedback on eBay.

utility that is distributed normally around the true store-level quality:

$$e_{\tilde{i}jst} \sim N(\xi_s, \widetilde{\sigma}_s^2) \tag{6}$$

where the subscript $\tilde{i}$ identifies consumers who purchased a good from seller $j$ before the arrival of consumer $i$. Upon receiving this signal, the consumer posts a rating, which is a transformation of the experience into a number which is in line with the rating system:

$$r_{\tilde{i}jst} = \rho e_{\tilde{i}jst}$$

where $\rho$ is a scaling term to transform the distribution of the signals, which is measured in utility, into the distribution of ratings. The platform displays the average of seller $s$'s ratings in the last six months as the rating score, $r_{st}$, along with the total number of ratings in the last six months, $n_{st}$.

Here, we assume that the posted ratings and, hence, the rating score are both continuous and have infinite support, something that is at odds with the actual setting. While we make these assumptions in order to simplify the updating procedure, we believe that this model serves as a good approximation of the process in this application. We provide further discussion of this model as an approximation in Appendix B.

Upon arrival, a consumer observes the current rating score, the number of ratings, and the seller's identity. Based on the seller's identity, the consumer believes that the seller's quality is drawn from a normal distribution:

$$\xi_s \sim N(\mu_s, \sigma_s^2) \tag{7}$$

which is the prior belief. The size of the variance in this distribution represents the level of uncertainty that the consumer has about the store's quality based purely on that store's identity. For example, a consumer may observe that the seller is walmart.com and be very certain about the seller's quality (i.e., a low value of $\sigma$), or she may observe a local seller and be uncertain about the quality (i.e., a high value of $\sigma$).

Then, using the average rating and number of ratings, the consumer updates her beliefs about the expected quality of the seller using Bayes' rule:

$$E[\xi_s|r_{st}, n_{st}] = \frac{\widetilde{\sigma}_s^2}{\widetilde{\sigma}_s^2 + n_{st}\sigma_s^2}\mu_s + \frac{n_{st}\sigma_s^2}{\widetilde{\sigma}_s^2 + n_{st}\sigma_s^2}\left(\frac{r_{st}}{\rho}\right) \tag{8}$$

This is a weighted average of the scaled rating score and the prior mean. The weights are based on the variance in the prior belief, the variance in the distribution of signals and the total number of signals which generated the rating score. Intuitively, the posterior is heavily weighted towards the prior when there are no ratings and converges to the scaled rating score as the number of ratings gets very large. Also note that because we assume that the experience signals are distributed around the true quality of seller $s$, the average rating, and thus, the belief of store quality, converges to the true store quality as the number of ratings gets large.

Importantly, we allow for *observed* heterogeneity in the learning process across sellers. That is, we specify the learning parameters as a function of the seller's type, $c_s$, and other seller covariates, $Y_s$. The mean of the prior and the ratio of the prior variance to the signal variance are written as function of these observables:

$$\mu_s = \mu_0 + \mu_N \mathbf{1}(c_s = N) + \mu_y Y_s \tag{9}$$

$$\frac{\sigma_s^2}{\tilde{\sigma}_2^2} = \bar{\sigma}_s^2 = exp\left(\sigma_0 + \sigma_N \mathbf{1}(c_s = N)\right) \tag{10}$$

where the vectors $\boldsymbol{\mu} = [\mu_0, \mu_N, \mu_y]$ and $\boldsymbol{\sigma} = [\sigma_0, \sigma_N]$ are parameters to be estimated. They each include a constant and a national specific shifter. In addition, we allow for the prior on quality to be a function of other seller level observables including seller tenure and a dummy indicating whether or not the seller sells more than one brand.[13]

The reason we parameterize the ratio of the variances rather than each of them separately is because we are not able to identify the ex ante variance of quality separately from the distribution of signals using only the demand model. As we discuss further in Section 4, we would need to use

[13] In other specifications, we included $Y$ in the sigma parameters as well, and the estimates of the learning parameters across seller type did not change significantly in these specifications.

either the sellers with zero ratings or the distribution of the individual ratings within a seller in order to separately identify these objects. However, the ratio is a sufficient measure of the learning process for our purposes, as it is a measure of the relative weight individuals place on the prior versus the rating score. A relatively large weight on the prior is either because of a relatively small prior variance or a relatively large signal variance. In either case, this large weight implies that the learning signals are not very informative, as either the consumers are confident in their ex ante belief on seller quality and/or the experience signals do not provide much information.

Because of the linearity in consumer preferences, the expected utility from buying product $j$ from seller $s$ is:

$$E[U_{ijst}] = \beta X_{jst} + \zeta_{st} + \epsilon_{ijst} = \beta X_{jst} + \frac{1}{1 + n_{st}\bar{\sigma}_s^2}\mu_s + \frac{n_{st}\bar{\sigma}_s^2}{1 + n_{st}\bar{\sigma}_s^2}(\frac{r_{st}}{\rho}) + \nu_{jst} + \epsilon_{ijst}$$

Given this information, consumer $i$ chooses to purchase the product/seller which gives her the highest expected utility.

There are a few of issues about this specification that deserve discussion. First, ideally, we would include some level of seller-level unobserved heterogeneity in the prior. While it is possible to include such heterogeneity via a seller fixed effect, it makes the model less tractable by adding complexity to the non-linear portion of the specification and also precludes us from estimating a type-specific effect in the prior. Instead, we allow for the unobserved heterogeneity to enter via a seller fixed effect in $X_{jst}$. This suggests that there is a portion of seller quality that is known to consumers ex ante (i.e., the fixed effect) and a portion that consumers learn about through seller-level observables and the ratings. However, we also have estimated a model where there is a seller fixed effect inside the prior, and the qualitative differences in the estimates of the sigma parameters across seller type are quite similar.

Next, note that we could assume that $\zeta_{st}$ (i.e., the learning process) is a linear function of ratings like in the reduced form analysis, which makes the model less complex. However, modeling it as a Bayesian process has a few advantages. First, in the linear model, the effect of the 'prior' is fixed in that it is not allowed to vary as the number of ratings changes. It is common in learning models to assume that the effect of the prior (or the 'weight' on the prior) may become smaller as

consumers receive more signals. Second, the linearity of the learning function is restrictive in that it doesn't allow for interpretation of coefficients to be weights placed on the signal(s) and the prior. Finally, similar to the above points, these specifications don't allow for an estimation of what are thought of as 'structural' learning parameters or the deep parameters of the learning process such as the prior beliefs of seller quality and the distribution of learning signals.

# 4 Estimation

As is standard in the discrete choice literature, we divide Equation 5 by the outside share and take logs to get:

$$log(s_{jst}) - log(s_{0t}) = \beta X_{jst} + \frac{1}{1 + n_{st}\bar{\sigma}_s^2}\mu_s + \frac{n_{st}\bar{\sigma}_s^2}{1 + n_{st}\bar{\sigma}_s^2}(\frac{r_{st}}{\rho}) + \nu_{jst} \tag{11}$$

We assume that $\nu_{jst}$ is independent of a set of instruments $Z_{jst}$, forming the following moment conditions:

$$E[\nu'_{jst} Z_{jst}] = 0$$

Given a vector of parameters, $\theta$, we calculate the empirical analogy of this condition:

$$M(\theta) = \frac{1}{n}\sum_n \nu'_{jst}(\theta) Z_{jst}$$

and form a GMM objective function:

$$G(\theta) = M(\theta)' W M(\theta)$$

where $W$ is a weighting matrix. The estimated parameters solve:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, G(\theta)$$

However, because there are a large number of parameters in $\beta$ and $\boldsymbol{\mu}$, we create an adjusted GMM problem that takes advantage of the linearity in the problem. The parameters are separated into the linear parameters, $\Omega = [\beta, \boldsymbol{\mu}, \rho]$, and the non-linear parameters in vector $\boldsymbol{\sigma}$. Given a value

of $\boldsymbol{\sigma}$, Equation 11 can be re-written as:

$$log(s_{jst}) - log(s_{0t}) = \beta X_{jst} + \omega_{st}(\boldsymbol{\sigma})\mu_s + (1 - \omega_{st}(\boldsymbol{\sigma}))(\frac{r_{st}}{\rho}) + \nu_{jst} \tag{12}$$

where:

$$\omega_{st} = \frac{1}{1 + n_{st}\bar{\boldsymbol{\sigma}}_s^2}$$

The problem is now transformed into a linear problem where the linear parameters can be estimated via a regression. This produces estimates of $\hat{\Omega}(\boldsymbol{\sigma})$ and $\hat{\epsilon}(\boldsymbol{\sigma})$. Then, we can form moment conditions based on the non-linear parameters:

$$E[\epsilon_{jst}(\boldsymbol{\sigma})' Z_{jst}^{(2)}] = 0$$

where $Z_{jst}^{(2)}$ is the subset of $Z_{jst}$ related to the non-linear parameters. The empirical analog of this is given by:

$$\tilde{M}(\boldsymbol{\sigma}) = \frac{1}{n}\sum_n \epsilon_{jst}(\boldsymbol{\sigma})' Z_{jst}^{(2)}$$

We form the adjusted GMM objective function:

$$\tilde{G}(\boldsymbol{\sigma}) = \tilde{M}(\boldsymbol{\sigma})'\tilde{W}\tilde{M}(\boldsymbol{\sigma})$$

and find the parameters that solve:

$$\hat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma}}{\operatorname{argmin}}\, \tilde{G}(\boldsymbol{\sigma})$$

Therefore, in order to estimate the model we follow the following procedure:

- Guess values of $\boldsymbol{\sigma}$.
- Estimate $\hat{\Omega}(\boldsymbol{\sigma})$ and $\hat{\epsilon}(\boldsymbol{\sigma})$ via an inner-loop regression.
- Form $\tilde{G}(\boldsymbol{\sigma})$
- Find $\boldsymbol{\sigma}$ that minimizes $\tilde{G}(\boldsymbol{\sigma})$

### Identification

The identification of the $\beta$ parameters and the $\rho$ parameter is straightforward given the formulation of the estimation procedure above. That is, these parameters appear on the observables in our inner-loop linear regression. One note is that prices are endogenous so, in practice, our inner loop regression is a two stage least squares regression. The instrument used for price is the total stock of tablets with the same specifications that other sellers are holding.[14]

The identification of the parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ is not as clear, but can be explained by looking at Equation 11. First, the parameters in $\boldsymbol{\sigma}$ are identified by variation in the market share resulting from variation in the interaction between the rating score and the number of ratings. This can be seen in the third term of Equation 11. Specifically, given the estimate of $\rho$, this term is a function of $r_{st}$, $n_{st}$, and $\boldsymbol{\sigma}$, implying that the relationship between the ratings (both the score and the number of ratings) and the market shares identify the parameters in $\boldsymbol{\sigma}$. This is intuitive as the elements of $\boldsymbol{\sigma}$ are parameters measuring how important ratings are in determining demand. It is then straightforward to see that the $\boldsymbol{\mu}$ parameters are then identified by the interaction between $n_{st}$ and the elements of $Y_s$. Intuitively, this is measured by how much the prior matters, or the elements of the prior, for a particular seller as consumers get more and more ratings.

Given the arguments above, the elements of $Z^{(2)}_{jst}$ we use are: (1) interactions between $r_{st}$, $n_{st}$, and each of the observables in $\sigma$ and (2) interactions between $n_{st}$ and each of the elements of sigma. Note that in order to identify the learning parameters, we assume that $r_{st}$ and $n_{st}$ are conditionally independent of the demand shocks, $\epsilon_{ijst}$, and product specific shock, $\nu_{jst}$. We include a seller-fixed effect in $X_{jst}$ in the main specification and in $\mu_s$ in a robustness check in order to account for any seller-level heterogeneity that is not captured in the observables. Therefore, the assumption is that the within-seller variation of $n_{st}$ and $r_{st}$ is independent of the demand shocks. This assumption would be violated if the demand shocks for sellers or products are correlated. Because we don't have the variation in the data or the instruments in order to address this type of endogeneity, we maintain the assumption that the seller fixed effects are able to capture variation

[14] First-stage estimates using a linear function of ratings suggest that this instrument is relevant. To argue its validity, we suggest that the inventory held by competitors is a measure of the upstream market conditions for this particular tablet and is not correlated with demand for this particular seller.

in unobserved seller attributes and the observed product characteristics are able to capture the variation in demand at the product level.

As previously mentioned, we do not identify the signal variances, $\tilde{\sigma}$, separately from the prior variance, $\sigma$. These parameters would be separately identified by utilizing information for the sellers who have zero ratings and/or using the distribution of ratings within an individual seller. We choose not to utilize these pieces of information because there are very few sellers in our data with zero ratings and because utilizing the within-seller distribution of ratings would require a richer model of how ratings are generated. Although a decomposition of $\bar{\sigma}$ is an interesting exercise, it is not necessary to accomplish the goals of this research. Specifically, $\bar{\sigma}$ is informative of the weight individuals place on the prior versus the signals, meaning it is a measure of how much people use the ratings in their decision-making.

# 5 Results

We present the results of the estimates in Tables 7 and 8 with standard errors in parentheses.[15] The standard errors were calculated numerically for the non-linear parameters. The preference parameter estimates suggest that people value screen size, storage and newer tablets. We find a negative effect of access to the internet via cellular network which likely comes from the fact that these tablets are not as popular as the versions that just have access to WiFi. The negative effect of memory is likely due to both the fact that there is little variation in this variable and the memory of tablets is much less salient than the size and storage. iPad's iOS is the most valued operating system, followed by Windows and then Android. The price coefficient is negative and significant.

We now turn to the learning parameters in Table 8. We present two sets of estimates. The first is our base model (specification (1)) and the one that coincides with the preference estimates in Table 7. In the second, we move the seller fixed effect inside of the prior mean and thus, eliminated the covariates.[16] Comparing the estimates, we see that the constant and the national dummy variable

[15]The seller/product combinations which had 0 sales are dropped from the regression (i.e., included the outside option). We explored correcting the market shares as in Gandhi et al (2013) and obtained similar results.

[16]In addition, we have estimated a model where the tenure of a seller was included as a covariate in $\boldsymbol{\sigma}$. The qualitative results on the learning variables by type are similar to the base model. Interestingly, we find that ratings matter more for sellers with a longer tenure on Tmall. It may be the case that the tenure variable captures a general

in $\boldsymbol{\sigma}$ do not vary much across specification, suggesting that the placement of the seller fixed effect does not impact the results. Because of this, we focus on specification (1) in the discussion that follows.

First, we note that the scaling term $\rho$ is estimated to be about 0.67. Using the price parameter and assuming a sufficient number of ratings, we can calculate that an increase of the rating score by 0.1 units represents an increase in willingness to pay of about \$70.

We now turn to the learning parameters. We find that an increase in seller tenure negatively affects the prior and that being a multi-brand seller positively affects the prior. The result on tenure is not necessarily in line with intuition, but it may be because it is proxying for something else such as the the natural decay of demand over time or the types of products sellers with a longer tenure carry.

Importantly, we find that national sellers have a higher prior than that of local sellers, implying that consumers have a belief that national sellers are higher quality, ex ante. This result could be for many reasons. For example, it could be that national sellers advertise more than the local sellers, have better customer service, have faster shipping, have a more efficient return system, etc.

At the same time, we find that the national dummy in the $\boldsymbol{\sigma}$ parameter is negative and significant. This means that people place less weight on the ratings for national sellers than they do for the local sellers, or that ratings are more influential for local sellers compared to national sellers. To see this, we calculated the weight placed on the ratings for national and local sellers over different values of $n_{st}$ and present it in Figure 5. We can see that the weight increases much faster for the local sellers: by the 1,000th rating it is nearly 1.0 for the local sellers and near 0.5 for the national sellers. This is consistent with the initial evidence presented in Section 2.

The fact that ratings are not as important for national sellers could be because the distribution of shopping experiences (or ratings) for national sellers is relatively wide and/or the ex ante belief about seller quality (i.e., the prior variance) is relatively tight. In order to provide some evidence for the source of this result, we perform an exercise with the raw data. For each seller, we calculate the standard deviation of their ratings, which serves as a measure of the dispersion of signals. We

time trend in the effect of ratings on the platform.

then compare the average of the dispersion across seller groups. For local sellers it is 0.62 and for national sellers it is 0.67, indicating that the signals are not as informative for national sellers. However, we cannot reject the hypothesis that these means are equal at the 5% level (t value of 1.61). So while we cannot rule out that consumers receive more informative signals and have a wider prior belief for local sellers, this provides suggestive evidence that the latter is a more important factor in determining the differences in the learning process.

Given that, one could speculate that a consumer's ex ante beliefs about a local seller is wider, or less certain, due to the fact that they do not have many prior experiences or interactions with these sellers, and therefore rely more on the online reputation system to guide them in their purchase decisions. For a national seller, on the other hand, consumers are fairly certain as to their quality (i.e., have a small prior variance), likely due to the fact that it is a seller that they know and have dealt with in the past. For example, a consumer may choose to put less emphasis on the reputation for the online stores of Target.com, BestBuy.com, and Wal-Mart.com while at the same time using the reputation system to help them determine whether or not a local seller is a reputable dealer.

Put together, these results imply two things about the role of ratings in shaping demand across seller types in this market. First, without ratings, a national seller will have higher demand than an equivalent local seller because national sellers are higher quality sellers on average. Second, because the weight on rating score is relatively low for national sellers, eliminating ratings will not substantially change the belief about these sellers' quality. This implies that national sellers who have high ex post quality will not lose a great deal of demand when ratings are removed, whereas local sellers who have high ex post quality will lose a lot, for example. So the relative prior primarily determines the direction of substitution between types, and the weighting parameter determines the relative magnitude of substitution between types when ratings are eliminated or changed.

### Ratings and Demand Across Seller Types

We use the results of the model to quantify the effect of online reputation on demand across the different seller-types. To do this, we 'eliminate' the rating information from a consumer's information set and force her to make her purchase decision based purely on her prior information.

In other words, she does not update her belief about seller quality using ratings and places all the weight on the prior mean for seller $s$. We fix all other aspects of the market, including who is on the platform (i.e., there is no exit or entry), the product mix of each seller, and the prices sellers are charging. This allows us to isolate the effect of reputation apart from these other market features.

Under this environment, we calculate the total market share across all eight months for each type of seller. The last row of Table 9 displays the market share with the ratings in columns 2 and 4 and the differences between the share with and without ratings in columns 3 and 5. The total market share for electronics retailers (i.e., the sum of local and national share) on Tmall drops from 86% to 71%. This suggests that the ratings are, in general, valuable to this part of the marketplace as it allows sellers to compete with the outside good. However, we see that they are mostly valuable to the local sellers: market share for these sellers drops from 66% to 31%, while it jumps from 20% to 39% for national sellers. Using the number of sellers of each type and the average price of a tablet, we estimate that a national seller gains about $550,000 and a local seller loses about $150,000 over the eight month period, on average. This represents significant overall gains and losses for these sellers.

This overall shift in demand from national to local sellers is a result of changes in demand for many heterogeneous sellers who vary in their value of ratings. The way ratings affect demand for an individual seller depends not only on which type they are, but also on how their rating score compares to their ex ante quality. In general, sellers can be separated into sellers whose ex post belief with ratings is higher than the ex ante belief (i.e., the prior for that type) and sellers whose ex post belief is lower than the ex ante belief. That is, sellers for which ratings increase the quality belief and sellers for which ratings decrease the quality belief. This suggests that the removal of ratings results in both within and across type substitution.

To take a first look at this heterogeneity, we display the effect of ratings for sellers of each type who are separated into quantiles based on their total sales in Table 9. All levels of local sellers lose demand when ratings are removed, while all levels of national sellers gain demand. However, the changes in demand are the largest for the biggest sellers of each type: the top 20% of national sellers more than double their demand and the top 20% of local sellers lose more than half of their

sales. The changes in demand fall as we move down the quantiles.

The best selling local sellers lose the most demand because these sellers can no longer differentiate themselves from the low quality local sellers, meaning they lose demand to their local counterparts and the national sellers. A similar problem exists for the best selling national sellers, but because the impact of ratings is small for them, the removal of ratings does not affect them to a great degree. In fact, they benefit from the removal of ratings because they are able to take demand from the high quality local sellers they were previously competing against. We note that, for these high volume sellers, the stakes are quite large with the gains and losses in the *millions* of dollars.

We examine the heterogeneity in the impact of ratings further in Table 10 where we present the effect of reputation across sellers of different quality levels. We don't observe true quality, but we can measure the relative ex post belief of quality with the observed rating score. In the table we display the total market share with and without ratings at different quality groups, where quality is measured by the average rating score at the end of the sample period. For example, national sellers with a 5.0 rating score at the end of our sample period had total of 0.07 market share with ratings and 0.08 market share without ratings. The important margin to focus on is the change in the group share when ratings are removed, rather than what happens to the share as quality increases or decreases. This is because the number of sellers varies across groups but does not vary within group.

For national sellers, we see that the removal of ratings helps all quality groups. Again, the primary reason these sellers benefit from the removal of ratings is because it means that the competitive, high quality local sellers can no longer differentiate themselves from the low quality sellers. The fact that both lowly rated and highly rated national sellers benefit is evidence that this mechanism outweighs any cost or benefit national sellers may get from hiding their true quality. This highlights the importance of the heterogeneity in the learning process, as homogenous learning parameters would result in highly rated sellers benefiting from the reputation system regardless of their type. Instead, we see that highly rated national sellers would prefer if there were no reputation mechanism.

For the local sellers, we see a different pattern. That is, the low rated sellers gain from the removal of ratings because their reputation is now hidden. This is beneficial because these sellers now look identical to the highly rated local sellers and, because ratings have an impact, they are able to take demand from them. Then, for the highly rated local sellers, demand decreases because they are no longer able to differentiate themselves from the lowly rated local sellers or the national sellers. Only the highest rated local sellers benefit, but this is likely due to correlation between their ratings and the other seller observables that are in the prior and/or the fact that there are very few sellers in this group.

In summary, the results of this exercise suggest that local sellers rely on the rating system in order to compete with national sellers on Tmall. Specifically, the reputation system allows high quality (i.e., highly rated) local sellers, who otherwise would not be differentiated from their low quality counterparts, to compete with the well-known, national sellers.

There are two important caveats to mention about the interpretation of the results. First, they are computed while keeping prices and other aspects of the market fixed. Intuitively, if we were to allow sellers to respond to the removal of reputation by adjusting their prices, the results would be dampened. That is, we expect that high quality local sellers to lower their price to capture some of the demand that is lost without the reputation signal. The extent to which the results change depend on consumer's price sensitivity and the marginal costs of sellers. The small price coefficient in Table 7 suggests that the ability for the local sellers to generate demand through price reductions is limited. While estimating a full supply-side model is out of the scope of this research, we have computed the same exercises above while allowing sellers to adjust their price via a reduced form pricing function, and the results do not change to a significant degree.[17]

Second, the model is estimated assuming that consumers consider all sellers and products available. We provide evidence in Section 3 that this does not bias the results. Importantly, we find

[17] We estimate the following linear pricing function:

$$log(P_{sjt}) = \beta X_{jt} + \gamma \zeta_{st} + \epsilon_{jst}$$

When removing reputation, we compute prices using the estimates of $\beta$, $\gamma$ and $\epsilon$, but replacing $\zeta_{st}$ with $\mu_s$. Importantly, this pricing function captures the fact that prices vary due to the beliefs held by consumers.

We have explored estimating a static Baysian Nash pricing model, but we believe that the assumption of static pricing is likely violated in this market as sellers manage their reputation (see Fan et al. (2016)) and estimating a dynamic pricing model is out of the scope of this research.

no evidence that the 'visibility' of sellers is different across seller type or that the effect of ratings changes when this assumption is relaxed. However, if the true model is one where consumer only consider a subset of more visible products/sellers regardless of seller type, then we would underestimate the impact of ratings for more visible sellers and overestimate the effect for less visible sellers. The reason for this is because the model would attribute a lack of a demand response for less visible sellers to a small impact of ratings, when in reality it is due to the fact that they are not in consumers' choice sets. Therefore, the fact that the most visible sellers are likely the ones with the highest sales suggests that the estimated decrease (increase) of demand for local (national) sellers when ratings are removed is somewhat conservative.

# 6 Conclusion

We studied the heterogenous effect of online reputation across sellers on Tmall, a large business-to-consumer platform in China. There are two primary results to highlight. First, the online reputation system on Tmall impacts sellers who are local or online only retailers more than it impacts sellers who are well-known through their national presence. This result complements the previous literature studying the impact of reputation across sellers who differ in their experience or brand affiliation.

Second, we find that the reputation system on Tmall is vital in shaping demand among heterogeneous sellers. Specifically, the results suggest that the rating score facilitates competition between sellers who are well known through their national presence and sellers who are less well known. The key mechanism at play is that the rating system allows high quality (i.e., highly rated) local sellers to differentiate themselves from low quality sellers on the platform.

Although we do not model this aspect specifically, the results imply that the rating system affects the mix and number of firms on the platform over the long run. To the extent that consumers and policy makers rely on an online platform like Tmall to be a marketplace with many different sellers, our results show that the online reputation system is important in determining this mix of firms. Further, and more generally, as 'mom and pop' shops try to compete with online giants like Amazon.com and Walmart.com, our results suggest that the online reputation systems in place are

vital in determining their success.

## References


Berry, S. (1994, Summer). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics 25*(2), 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995, July). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

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

Chernew, M., G. Gowrisankaran, and D. P. Scanlon (2008). Learning and the value of information: Evidence from health plan report cards. *Journal of Econometrics 144*(1), 156–174.

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

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science 49*(10), 1407–1424.

Elfenbein, D. W., R. Fisman, and B. McManus (2015). Market structure, reputation, and the value of quality certification. *American Economic Journal: Microeconomics 7*(4), 83–108.

Fan, Y., J. Ju, and M. Xiao (2016). Reputation premium and reputation management: Evidence from the largest e-commerce platform in china. *International Journal of Industrial Organization 46*, 63–76.

Fang, L. (2018). Evaluating the effects of online review platforms onrestaurant revenues, consumer learning and welfare. *Working paper*.

Grennan, M. and R. Town (2015). Regulating innovation with uncertain quality: Information, risk, and access in medical devices. NBER working paper 20981.

Hollenbeck, B. (2016). Online reputation mechanisms and the decreasing value of brands. Working Paper.

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp.com. Harvard Business School Working Paper, No. 12-016.

Newberry, P. W. (2016). An empirical study of observational learning. *The RAND Journal of Economics 47*(2), 394–432.

Nosko, C. and S. Tadelis (2015). The limits of reputation in platform markets: An empirical analysis and field experiment. NBER working paper 20830.

Saeedi, M. and N. Sundaresan (2016, March). The value of feedback: An analysis of the reputation system. Working Paper.

Su, L., P. Xu, and H. Ju (2016). Common threshold in quantile regressions with an application to pricing for reputation. *Econometric Reviews*. Forthcoming.

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

Zhu, F. and X. Zhang (2006). The influence of online consumer reviews on the demand for experience goods: The case of video games. *ICIS 2006 Proceedings*, 25.

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

# A Tables and Figures

## A.1 Tables

Table 1: Information for Top 5 Sellers of Each Group

| Group | Rank | # of offline stores | # of employees | Website |
|---|---|---|---|---|
| 1 | 1 | 22 | . | Yes |
| 1 | 2 | 177 | . | Yes |
| 1 | 3 | 838 | . | Yes |
| 1 | 4 | 33 | . | Yes |
| 1 | 5 | 900 | . | Yes |
| 2 | 1 | . | 200-499 | No |
| 2 | 2 | . | 1-50 | No |
| 2 | 3 | . | 50-99 | No |
| 2 | 4 | . | 20-99 | No |
| 2 | 5 | . | 1-49 | No |
| 3 | 1 | . | 1-49 | No |
| 3 | 2 | . | 1-49 | No |
| 3 | 3 | . | 20-50 | No |
| 3 | 4 | . | . | No |
| 3 | 5 | . | 10 | No |

Notes: Displayed are the number and employees for the top 5 seller of each type in terms of sales. This information was gathered by searching for each seller's official website where, often times, the number of offline retail outlets are displayed. When the seller did not have an official website, we search for the store in a business directory that lists the number of employees. Note that a '.' indicates that the information was not available for that seller.

Table 2: Description of the market

| | All | | Electronics Only | |
|---|---|---|---|---|
| | Total | Average per month | Total | Average per month |
| No of sellers | 670 | 461 | 562 | 391 |
| Sales | 771,450 | 94,972 | 670,919 | 83,222 |
| Revenue($\$000,000$) | 192.7 | 23.4 | 164.0 | 20.1 |
| No of brands | 87 | 74 | 85 | 69 |
| No of models | 1,245 | 697 | 1,211 | 673 |

Notes: Displayed are the aggregate statistics for sellers of tablets on Tmall both over the entire 8 month sample period and the average per month. Prices are converted to US dollars using an exchange rate equal to 6.33.

Table 3: Description of Electronics Retailers by Type

| | Mean | Median | StDev |
|---|---|---|---|
| **National Retailers** | | | |
| Sales | 2,125 | 171 | 6,572 |
| Revenue($) | 444,840 | 36,768 | 1,218,893 |
| Price($) | 270 | 205 | 272 |
| No. of Brands | 1.20 | 1.00 | 0.83 |
| No. of Models | 7.13 | 5.00 | 6.35 |
| Tenure (months) | 33 | 31 | 21.6 |
| **Local Retailers** | | | |
| Sales | 1,188 | 65 | 4,438 |
| Revenue($) | 318,111 | 17,229 | 1,697,531 |
| Price($) | 289 | 215 | 272 |
| No. of Brands | 1.27 | 1.00 | 0.68 |
| No. of Models | 7.37 | 4.00 | 8.44 |
| Tenure (months) | 27 | 26 | 18.22 |

Notes: Displayed are seller-level statistics separated by seller-type. All statistics are averages across stores for the entire 8 month time period except for price, which are averages across products, stores, and months. Prices are converted to US dollars using an exchange rate equal to 6.33.

Table 4: Average Characteristics of Tablets

| | National | Local | All |
|---|---|---|---|
| Top Brands | | | |
| Teclast(#1) | 0.30 | 0.16 | 0.19 |
| Onda(#2) | 0.22 | 0.10 | 0.13 |
| Cube(#3) | 0.00 | 0.16 | 0.13 |
| Apple(#4) | 0.01 | 0.13 | 0.10 |
| Sumsung(#9) | 0.00 | 0.05 | 0.04 |
| Operating System | | | |
| Android | 0.60 | 0.61 | 0.61 |
| IOS | 0.01 | 0.13 | 0.10 |
| Windows | 0.46 | 0.30 | 0.34 |
| Cellular Internet Access | 0.27 | 0.26 | 0.26 |
| Screen Size(Inches) | 8.64 | 8.46 | 8.50 |
| Memory(GB) | 1.76 | 1.93 | 1.91 |
| Storage(GB) | 29.86 | 27.40 | 27.75 |
| Product Age(years) | 0.43 | 0.65 | 0.60 |
| Observations | 2,051 | 12,241 | 14,292 |

Notes: Displayed are the average characteristics for tablets sold by electronics retailers. Tablets may feature more than one operating system resulting.

Table 5: Ratings and Demand by Type

| | National | Local |
|---|---|---|
| Total Share | 19.57% | 65.82% |
| Mean Share | 0.30% | 0.20% |
| StDev Share | 1.00% | 0.70% |
| | | |
| Mean # of Ratings | 7447 | 4411 |
| Median # of Ratings | 1343 | 843 |
| StDev # of Ratings | 25075 | 15756 |
| | | |
| Mean Rating Score | 4.70 | 4.72 |
| Median Rating Score | 4.74 | 4.77 |
| StDev Rating Score | 0.33 | 0.26 |

Notes: Displayed in the first row is the average monthly market share of national and local sellers across our sample period. The remainder of the statistics are calculated using observations at the seller/month level.

Table 6: K-S Test of Equal Distribution of Ratings

| Test | Test Statistic | p-val |
|---|---|---|
| K-S Test of Equal Distributions | 0.1568 | 0.000 |
| K-S Test of National $\leq$ Local | 0.1568 | 0.000 |

Notes: The table displayed results of Kolmogorov-Smirnov tests. The null-hypotheses of the first row are the distributions are equal and the null-hypotheses for the second row is that the distribution of ratings for national sellers is lower than that of local sellers.

Table 7: First Stage Estimation Results

| Variable | Coefficient | Variable | Coefficient |
|---|---|---|---|
| Price | -0.001*** | Android | 0.018*** |
| | (0.6e-5) | | (0.7e-5) |
| Screen Size(Inches) | 0.016*** | IOS | 6.122*** |
| | (0.2e-5) | | (0.2e-4) |
| Storage(GB) | 0.008*** | Windows | 0.210*** |
| | (0.1e-6) | | (0.8e-5) |
| Ram | -0.012*** | Product Age | -0.839*** |
| | (0.3e-6) | | (0.2e-5) |
| Cellular Internet Access | -0.321*** | Constant | -13.876*** |
| | (0.3e-5) | | (0.001) |
| Top 19 Brand Dummies | Yes | Obs. | 12,348 |

Notes: Standard errors in parentheses, $^{*}(p < 0.05)$,$^{**}(p < 0.01)$,$^{***}(p < 0.001)$ Displayed are the results of the first-stage estimation of preference parameters. Product ages are measured in years, and normalized to $[0, 2]$.

Table 8: Second Stage Estimation Results

| Specification | | Linear FE | FE in $\mu$ |
|---|---|---|---|
| $\mu_{st}$ | Constant | 6.841*** | |
| | | (0.001) | |
| | National | 1.386*** | |
| | | (0.001) | |
| | Tenure | -2.186*** | |
| | | (0.001) | |
| | Multi-Brand | 0.551*** | |
| | | (0.2e-4) | |
| $\bar{\sigma}_{st}$ | Constant | -3.644*** | -4.563*** |
| | | (0.001) | (0.068) |
| | National | -3.253*** | -6.292*** |
| | | (0.002) | (0.005) |
| $\rho$ | | 0.671*** | 1.440*** |
| | | (0.1e-3) | ( (0.1e-3) |
| Observations | | 12,348 | 12,348 |

Notes: Standard errors in parentheses, $^{*}(p < 0.05)$,$^{**}(p < 0.01)$,$^{***}(p < 0.001)$ Displayed are the results of the second-stage estimation of the learning parameters. Observations without inventory information are dropped.

Table 9: The Effect of Ratings on Demand

| Market | National | | Local | |
|---|---|---|---|---|
| Share | w. Ratings | Remove Ratings | w. Ratings | Remove Ratings |
| Quantile | Market Share | $\Delta$ Market Share | Market Share | $\Delta$ Market Share |
| 1 | 17.39 | 18.38 | 61.25 | -33.79 |
| 2 | 1.78 | 1.47 | 3.59 | -1.14 |
| 3 | 0.33 | 0.21 | 0.80 | -0.17 |
| 4 | 0.07 | 0.06 | 0.16 | -0.03 |
| 5 | 0.01 | 0.00 | 0.03 | -0.01 |
| Total | 19.57 | 20.13 | 65.82 | -35.14 |

Note: The last row of the table table shows the change in type level market shares during September, 2014 to April, 2015 when ratings are removed. The remaining rows show the changes in market shares for the sellers of each type who are binned by their sales over the entire sample. Each bin contains 20% of the sellers.

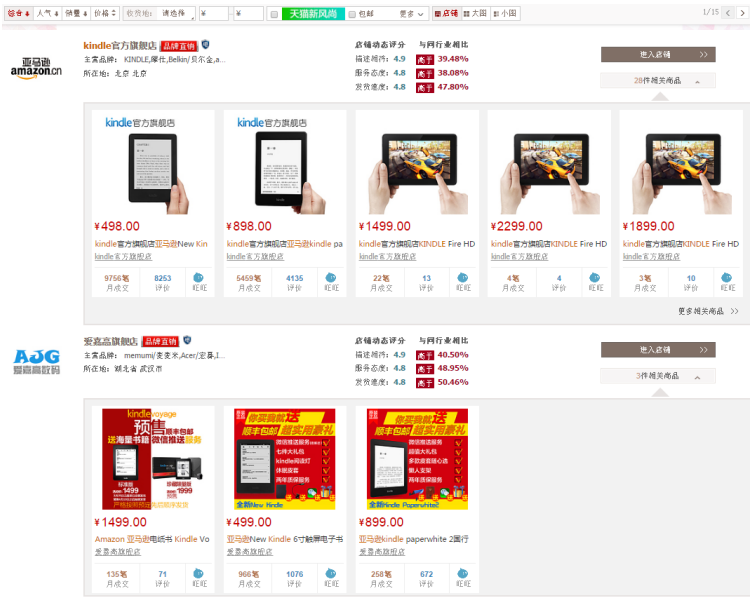
Table 10: The Effect of Ratings on Demand by Quality Group

| Rating | National | | Local | |
|---|---|---|---|---|
| | w. Ratings Market Share | Remove Ratings Δ Market Share | w. Ratings Market Share | Remove Ratings Δ Market Share |
| 4.3 | - | - | 0.02 | 0.02 |
| 4.4 | - | - | 0.02 | 0.03 |
| 4.5 | 0.04 | 0.04 | 0.11 | 0.06 |
| 4.6 | 0.65 | 0.14 | 4.13 | -2.33 |
| 4.7 | 15.39 | 18.29 | 31.18 | -18.45 |
| 4.8 | 3.19 | 1.89 | 28.32 | -14.42 |
| 4.9 | 0.44 | 0.59 | 2.40 | -0.43 |
| 5.0 | 0.07 | 0.08 | 0.21 | 0.16 |
| Total | 19.57 | 20.12 | 65.82 | -35.14 |

Note: The last row of the table table shows the change in type level market shares during September, 2014 to April, 2015 when ratings are removed. The remaining rows show the changes in market shares for the sellers who are in binned by their rating score at the end of the sample period.

## A.2 Figures

Figure 1: Rating Information on the Search Page

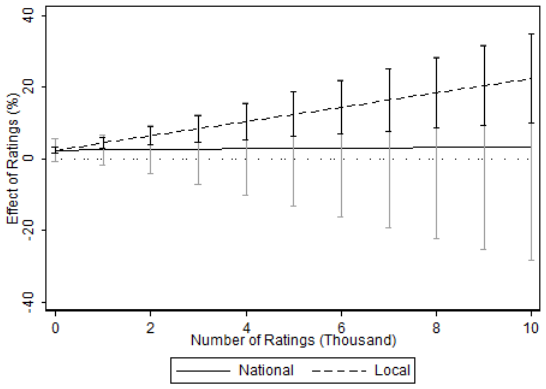



Notes: Displayed is a screen shot of a search page after a consumer sorts the default search by seller. A seller's average ratings are displayed at the top, while the number of ratings for each product are displayed below the products price and description.

Figure 2: Ratings Information on the Seller Page


Notes: Displayed is a screen shot of seller's home page. The average and the distribution of ratings are prominently displayed and the number of ratings appear above the distribution.

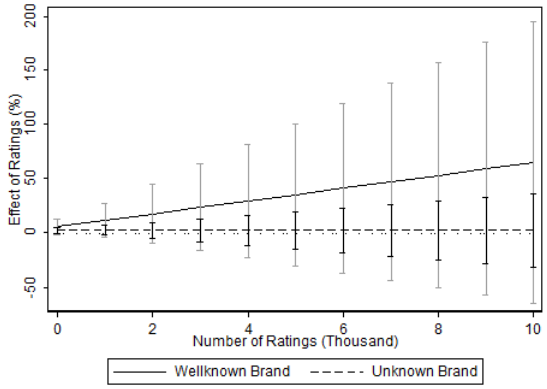
Figure 3: The Effect of Ratings on Sales



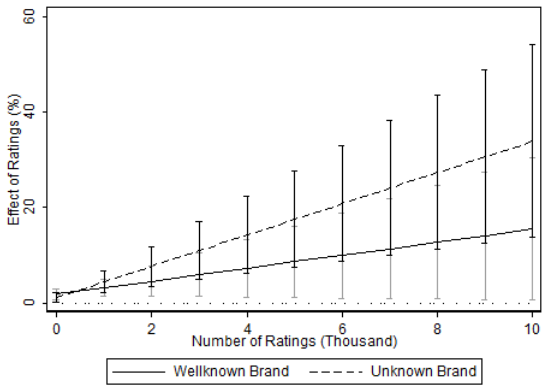

Note: The figure presents the estimated effect of ratings on sales across the two seller types. The y-axis is the percentage change in quantity associated with an increase in rating score of 0.01 at the value of $n$ indicated on the x-axis.

Figure 4: The Effect of Ratings on Sales Across Brands

(a) National Sellers

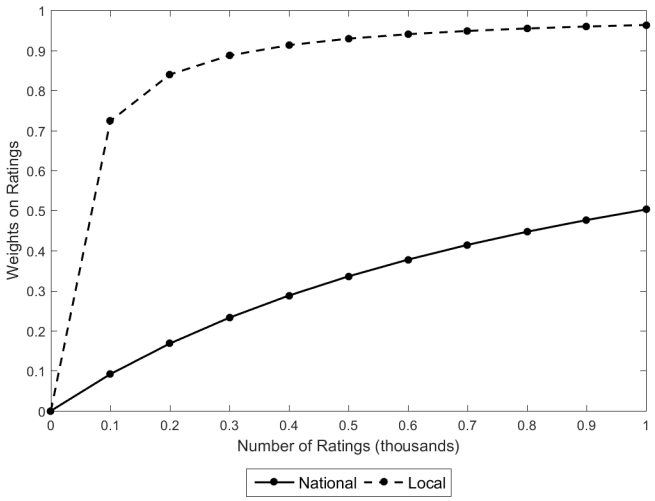



(b) Local Sellers




Note: The figures present the estimated effect of ratings on sales across the two seller types and across brands. The y-axis is the percentage change in quantity associated with an increase in rating score of 0.01 at the value of $n$ indicated on the x-axis. A well-known brand is one that is in the top-5 of sales for our sample period.

Figure 5: Weight on the Ratings by Seller Type




Notes: The figure displays the weight placed on the rating score over different values of $n$ for each seller type. These are calculated using the estimates of the model.

# B Rating Model

The primary reason for assuming that ratings come from a continuous and infinitive support is that it simplifies the updating procedure substantially. That is, suppose we assumed a model of how ratings are generated in which a consumer receives the experience signal, $e_{\tilde{ijst}}$, and then posts a rating based on a series of thresholds:

$$r_{\tilde{ijst}} = \begin{cases} 1, & \text{if } \rho e_{\tilde{ijst}} \in [-\infty, G_1) \\ 2, & \text{if } \rho e_{\tilde{ijst}} \in [G_1, G_2) \\ 3, & \text{if } \rho e_{\tilde{ijst}} \in [G_2, G_3) \\ 4, & \text{if } \rho e_{\tilde{ijst}} \in [G_3, G_4) \\ 5, & \text{if } \rho e_{\tilde{ijst}} \in [G_4, \infty) \end{cases} \tag{13}$$

After observing the average rating and the number of ratings, consumer $i$ calculates the expectation of the $e_{\tilde{ijst}}$ which replaces $(\frac{r_{st}}{\rho})$ in Equation 8. If the consumer observes the total number of ratings and not the individual ratings themselves, then this calculation becomes computationally intractable. For example, suppose the consumer observes $n_{st} = 100$ and $r_{st} = 4.5$. She would have to integrate over all the possible combinations of the 100 ratings that would lead to an average rating score of 4.5 in order to calculate the expected value of the signals. Considering the fact that most sellers have $n_{st}$ in the thousands, calculating beliefs under this of assumption is infeasible. Under our assumption of a continuous and infinite distribution of ratings, this calculation is not necessary.

Another option is to use the discrete model and assume that consumer $i$ observes the complete distribution of ratings, something that is observable on the seller's page, which simplifies the problem substantially. That is, the expectation of the signal becomes:

$$E[e_{\tilde{ijst}}|\{n_{cst}\}, g \in [1,2,3,4,5]] = \sum_{g \in [1,2,3,4,5]} [e_{\tilde{ijst}}|\rho e_{\tilde{ijst}} + y \in [G_{m-1}, G_m)] \times \frac{n_{cst}}{n_{st}}$$

where $n_{cst}$ indicates the number of ratings which are equal to $g$. This is a straightforward calculation

because the assumed distribution of signals is normal. We have estimated the model under this structure, with little qualitative change in results.

Additionally, we believe that our model serves as a reasonable approximation of belief formation in this market for two reasons. First, the scale parameter, which will be estimated, serves to adjust the variance of ratings such that a realization of a rating outside of the range of 1 to 5 occurs with low probability. Second, with only a small number of ratings (e.g., 100), the average rating from the discrete model above approximates the average rating from the continuous model quite closely. The median seller in our sample has over 800 ratings.