Association for Information Systems

AIS Electronic Library (AISeL)

ECIS 2022 Research Papers

ECIS 2022 Proceedings

6-18-2022

Does Platform Competition Drive Ratings Inflation? The Impact of Vertical Spillover Effects

Yulia Vorotyntseva
Saint Louis University, yulia.vorotyntseva@slu.edu

Aleksi Aaltonen
Temple University, aleksi@temple.edu

Subodha Kumar
Temple University, subodha@temple.edu

Paul Pavlou
The University of Houston, pavlou@bauer.uh.edu

Follow this and additional works at: https://aisel.aisnet.org/ecis2022_rp

Recommended Citation

Vorotyntseva, Yulia; Aaltonen, Aleksi; Kumar, Subodha; and Pavlou, Paul, "Does Platform Competition Drive Ratings Inflation? The Impact of Vertical Spillover Effects" (2022). *ECIS 2022 Research Papers*. 83.
https://aisel.aisnet.org/ecis2022_rp/83

This material is brought to you by the ECIS 2022 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2022 Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

DOES PLATFORM COMPETITION DRIVE RATINGS INFLATION? THE IMPACT OF VERTICAL SPILLOVER EFFECTS

Research Paper

Yulia Vorotyntseva, Saint Louis University, yulia.vorotyntseva@slu.edu

Aleksi Aaltonen, Temple University, aleksi@temple.edu

Subodha Kumar, Temple University, subodha@temple.edu

Paul Pavlou, University of Houston, pavlou@bauer.uh.edu

Abstract

The familiar 5-star ratings system is an important information source for consumers deciding where to eat or what products to buy. Ideally, a retail platform owner should safeguard ratings against various biases, yet platform owners sometimes let average ratings become inflated. We study a situation in which a platform faces competition from another platform that offer the same items and, consequently, consumers may see different ratings for items across the platforms. Using a series of experiments in an online food ordering setting, we show that consumers are more likely to buy the item from a platform where it is rated higher. Therefore, a platform that offers lower but perhaps more accurate ratings risks hurting itself by not letting its ratings become inflated. We explain this by a vertical spillover effect by which diverging ratings across platforms influence platform choice and discuss implications to platform owners, regulators, and consumers.

Keywords: Experiment, Ratings, Spillover effect.

1 Introduction

The importance of online ratings has steadily grown as a means by which consumers decide where to eat, what products to buy, or which doctor to visit (Ibbotson 2018, Sahoo et al. 2018). Platform vendors are aware of this and typically develop a host of legitimate, semi-legitimate, and outright fraudulent tactics to boost their ratings, to which platform owners respond with various countermeasures. However, at the same time, the platform owners sometimes let the average ratings on their platform to become inflated. In particular, there is a risk that intensifying competition between platforms could motivate platform owners to collude with their vendors and to turn a blind eye to attempts to artificially boost the ratings (Financial Times 2018, Wang et al. 2020). This could happen if the inflated ratings could give a platfom an advantage over other platforms at the cost of reducing the informativeness of the average ratings to consumers. Ultimately, ratings inflation may destroy the value of average ratings as a source of information to consumers (Filippas et al. 2018).

Retail platforms typically adopt a familiar 5-star ratings system as a way of collecting and displaying user-generated product evaluations, which makes it easy for consumers to use product evaluations across competing platforms. Chevalier and Mayzlin (2006, p. 345) note that *”there is nothing to stop a consumer from using the information provided by a one Web site to inform purchases made elsewhere”* and, for instance, 15.7 percent of our experimental subjects said that they check items on multiple platforms before making a purchase decision. Consumers may thus often compare average ratings for the same items across competing platforms (Gao et al. 2015, de Langhe et al. 2016), making salient the fact that average ratings for the same items diverge across platforms (Chevalier and Mayzlin 2006, Zervas et al. 2020). This raises an important question what kind of platforms may benefit or lose from such shopping around for information by consumers.

There is some evidence that platform owners may be inclided to react strategically to consumers' propensity to multihome in their information seeking behavior. Kathuria and Lai (2018) suggest that an accumulated body of ratings and reviews may give a platform an advantage over its competitors, whereas Sahni (2016) observe in a slightly different context that the strategic benefits of making consumers more informed do not necessarily accrue to the focal actor due to spillover effects. However, we do not know if there is a real threat from hosting lower average ratings than competing platforms or whether such a threat is merely an unfounded perception. To this end, we draw from information systems and marketing literature that have analyzed spillover effects from consumer perceptions of a product, service, or a company to the sales of nearby entities in the same category (e.g. Borah and Tellis 2016, Janakiraman et al. 2009, Kumar et al. 2018b, Roehm and Tybout 2006). In contrast to much studied *horizontal spillover effects* between entities that occupy the same position in the value chain, we hypothesize that the way in which digital retail platforms integrate items and, for instance, their delivery into a seamless consumer experience can give rise to a *vertical spillover effect* that affects platform choice.

Vertical spillover effect suggests that consumers prefer to order from a platform where an item is rated higher even if they will get the same item regardless of the platform choice. This may happen when the consumer is shopping for a specific product or service and checks it on multiple retail platforms to gather more information, or in a less obvious manner, when a consumer would first encounter an item on one platform and forego the purchase due to its low rating, but later sees the item on another platform where it is rated higher and now proceeds with the purchase. To study the existence of vertical spillover effect in platform choice, we ask the following research question: *Do diverging ratings for the same item across competing platforms influence the choice of the platform used to buy the item?* The answer has important managerial implications in the context of platform competition when both vendors and consumers multihome, that is, use simultaneously multiple platforms. The presence of a vertical spillover effect would also mean that some sort of collective (regulatory) intervention may be needed to maintain the integrity of online ratings systems in the long run. The direct effect of vertical spillovers would seem innocuous to consumers who nevertheless end up purchasing the same item. However, ratings inflation can degrade the value of average ratings over time as a source of information and thus lead to increasing purchase errors (Filippas et al. 2018).

We conduct an experimental study in a restaurant food delivery setting using a combination of incentive aligned behavioral experiments and experiments based on stated preferences method. In each experiment, subjects choose simultaneously a restaurant and a platform from which they want to place a food delivery order. For some of the restaurants, the average rating varies between platforms whereas for others the average rating is the same between the platforms. Importantly, we design the experiments so that they account for different platform attributes and possible *a priori* preference for one or the other platform among the subjects. For the platforms, we choose Yelp and GrubHub that both operate in the US and share the same delivery system, and therefore the subjects can expect to receive exactly the same food and delivery experience regardless of the platform they choose. In other words, the experiments are constructed to identify the impact of diverging average ratings for the same item across platforms on platform choice. We also conduct a supplementary study in which we look at the possible moderating effect of the number of ratings on the vertical spillover effect.

The results show that platform choice is affected by diverging average ratings across platforms—more consumers choose the platform where the chosen restaurant is rated higher even when it is strongly implied that there will be no difference between the quality of food or overall consumer experience regardless of the platform choice. This is consistent with the presence of a vertical spillover effect and suggests that it may be disadvantageous for a platform to counter ratings inflation if competing platforms do not reciprocate such actions. Consequently, platform owners need to manage the 5-star system carefully to support consumer decision making while avoiding strategic mistakes in competition with other platforms. Interestingly, we find no evidence of a moderation effect from the number of ratings. The subjects appear to ignore the number of ratings and be guided by the average rating even when it is based on only one data point. Finally, the results of the stated preferences study are remarkably consistent with the incentive aligned experiments, although the data is somewhat noisier in the former case.

2 Literature Review

2.1 Online Ratings

Our work draws from several streams of research into online ratings and reviews. These are concerned with the impact of ratings and reviews on purchase decisions and sales (Chevalier et al. 2018, de Langhe et al. 2016, Forman et al. 2008, Li 2018, Rocklage and Fazio 2020, Sun 2012, Watson et al. 2018, Yin et al. 2016), reviewer motivation and behavior (Moe and Schweidel 2012, Pagano and Maalej 2013, Shen et al. 2015), the design of ratings systems (Chen et al. 2018, Hu et al. 2009, Huang et al. 2019, Jiang and Guo 2015, Lee and Kai 2020, Tunc et al.) and, in particular, spillover effects from ratings and reviews to other products and sellers (Jabr and Zheng 2014, Kumar et al. 2018a, Sahni 2016). Studies show that average ratings can reflect the quality of purchased items and the overall consumption experience remarkably well (Gao et al. 2015), but also that the ratings include various biases and are influenced by factors unrelated to the quality of the purchased items or the overall consumption experience. Such factors include the heterogeneity of platforms in terms of consumer tastes (Zimmermann et al. 2018), fake and strategic reviewing behavior (Ho et al. 2017, Kumar et al. 2019, Sahoo et al. 2018, Wang et al. 2020), managerial interventions by vendors (Ananthakrishnan et al. 2020, Kumar et al. 2018a,b), new reviewers being primed by previous reviews (Godes and Silva 2012), and the accuracy of product information provided on the platform. As a result, the same item is often rated differently on different platforms, which may be further amplified by platform-specific sequential and temporal dynamics that affect the body of ratings over time (Godes and Silva 2012, Lee et al. 2015b, Moe and Schweidel 2012).

Among the potential biases that affect online ratings, there is increasing evidence of ratings inflation. Average ratings tend to increase over time and lose their variance as they become compressed toward the maximum value (Athey et al. 2019, Filippas et al. 2018, Hu et al. 2009, Kokkodis 2021, Nosko and Tadelis 2015, Zervas et al. 2020). The inflation of average ratings can increase sales in the short term, but it also threatens to render the 5-star system uninformative for consumers if at some point ratings stop effectively differentiating the items (Aziz et al. 2020, Kokkodis 2021). It is also concerning that consumers sometimes show overrealiance with respect to user-generated ratings (Aziz et al. 2020, Kokkodis 2021). Literature identifies different reasons for ratings inflation, which generally relate to reviewer and vendor behavior that 'push' average to become inflated (Fradkin et al. 2015; Hu 2009; Lee and Kai 2020). However, no study has explained platform owners' 'pull' toward higher average ratings in the context of platform competition, which is the gap we address in this paper.

2.2 Platform Competition

Platform competition does not always follow a simple winner-take-all logic but instead platforms are heterogeneous and try differentiate from competition along several dimensions (Huotari et al. 2017, Rietveld and Schilling 2020, Schilling 2002). For instance, food delivery, ride sharing, retail and lodging are industries where there are multiple competing platforms that try protect their market share while paying close attention to what their competitors do. In addition to strategic maneuvering by platform owners, platform competition is shaped by cognitive biases among consumers (Katsamakas and Madany 2019), multihoming behavior on different sides of the platform (Kim et al. 2017), and spillover effects (Krijestorac et al. 2020), which all make competitive dynamics more complex than the winner-takes-all scenario based on strong network effects suggests. We find that the 5-star system has received little attention in this context despite its impact on consumer behavior.

Zervas et al. (2020) show that cross-listed properties have typically higher ratings on AirBnB than on TripAdvisor, further confirming our observations about the presence of diverging ratings for the same items across platforms. Chevalier and Mayzlin (2006) analyze the impact of diverging ratings on sales and find that a higher rating for a book on a platform results in more sales on the platform compared to its competitor with a lower rating for the same book. The authors do not study whether demand actually shifts from a platform to another or whether consumers make a choice between the platforms based on the ratings, yet they note that *"there is nothing to stop a consumer from using the information provided by one Web site to inform purchases made elsewhere"* (Chevalier and Mayzlin 2006, p. 345). At the same time, papers on fake reviews have shown that misinformation can have short-term benefits from a platform owner's perspective (Ananthakrishnan et al. 2020, Wang et al. 2020). Kathuria and Lai (2018) identify online ratings and reviews as a strategic issue in platform competition and discuss their portability across platforms from a legal perspective. The authors argue that a dominant platform tends to have more reviews and, *"other things being equal, users will prefer a platform that has a larger number of reviews"* (p. 1294). At the same time, other studies have found mixed evidence on the impact of review volume on sales (Blal and Sturman 2014, Watson et al. 2018, Zimmermann et al. 2018).

3 Spillover Effects

Extant studies have not identified exact mechanisms by which platform-specific bodies of ratings could influence platform competition and, thus, whether competition motivates platform owners to let ratings become inflated. We argue that this may happen due to a spillover effect. Spillover effects are defined as externalities by which an event in one context influences another event in a proximate but essentially unrelated context (Xu and Schwarz 2018). Marketing research has studied spillover effects from product, company, or brand-related events to neighboring entities in the same category (e.g. Borah and Tellis 2016, Janakiraman et al. 2009, Roehm and Tybout 2006). Spillovers have also been found to exist in the context of online ratings and reviews. Sahni (2016) finds that highly rated competitors may reap a substantial share of benefits from advertising due a spillover effect, up to a point that advertising may hurt the advertiser under certain conditions. Jabr and Zheng (2014) study spillover effects from the reviews of competing products to a focal product, which is extended by Pavlou et al. (forthcoming) who take a market basket approach to study how perceptions from the reviews of co-visited products spill over to the purchase decision of the focal product. Kumar et al. (2018a) investigate a spillover effect from management responses to the reviews of company products to the competitors of the company. We call these and similar spillovers as horizontal spillover effects as they take place between entities that occupy the same position in the value chain.

The way in which retail platforms integrate different parts of the value chain into a unified consumer experience can give rise to spillover effects between entities that occupy different positions in the value chain. Such vertical spillovers have been studied relatively little, yet they could explain how diverging ratings of the same item across platforms affect platform choice (competition). Li and Agarwal (2017) show that large third-party developers benefit from a positive spillover effect from Facebook's tight integration of Instagram as a first-party app (whereas small developers experience a negative spillover effect), but seemingly no other paper has studied vertical spillovers in the platform context. At the same tine, it is intuitive that whenever a retail platform lists a huge number of items and their average ratings, consumers may not be able to attribute the average ratings sharply to the items only, but the ratings for individual items may taint the evaluation of the platform itself. For instance, Nosko and Tadelis (2015) find that consumers *"draw conclusions about the quality of the platform from single transactions, causing a reputational externality across sellers"* (Nosko and Tadelis 2015).

To summarize, a vertical spillover effect would mean that consumers are not only more likely to buy items that have higher ratings—which is well known by the literature—but also that they are more likely to choose to buy from a platform where an item is rated higher as compared to an alternative platform where the same item is available. This can alter platform competition in contexts where the same vendors and consumers are present (multihome) on multiple competing platforms.

3.1 Investigating the Vertical Spillover Effect

To investigate the existence of a vertical spillover effect, we study a restaurant food delivery setting that offers a good example of a complex service system behind a uniform platform front end: restaurants prepare meals that are delivered by companies such as Deliveroo, DoorDash or GrubHub, while the two (meal and delivery) are bundled together and sold by a platform such as Waiter.com, Yelp or, again, GrubHub that may collect their own ratings and reviews or source these from another platform. We assume that vertical spillovers emerge as follows. To order a meal, a consumer must choose between different types of meals and from which restaurant to order. This activates a behavioral mind-set that makes comparative procedures cognitively highly salient to the consumer (Xu and Schwarz 2018). If the consumer then retains the same mind-set with respect to a platform choice (as one could expect), the consumer may fall back on and confuse (now) a false signal from diverging average ratings across platforms for the chosen item as an indication of a quality difference between the platforms. This is also known as a ’halo effect’ that results from the inability of humans to form fully separate impressions of different parts of a whole (Borah and Tellis 2016).

4 Methodology

We conduct a series of experiments using both incentive aligned and stated preferences method in an online food delivery setting. In ‘incentive aligned’ experiments the subjects are motivated to choose according their true preferencs, since the experiments are designed so that the payoffs depend on the choices that the subjects make, whereas ‘stated preferences method’ means that we ask the subjects to imagine a scenario and make choices as if the scenario was real. The food delivery setting is particularly suitable for the incentive aligned setup because (1) it allows us to set a uniform monetary value on subjects’ rewards, (2) it provides subjects with substantial flexibility of choosing a meal they like and, hence, the rewards are likely to be desirable for the subjects, and (3) we can ensure that the subjects cannot transfer their rewards or return them to the vendor in exchange for cash.

We choose Yelp and GrubHub that are two well known food delivery platforms in the US. Both platforms use the same GrubHub delivery system, which means that consumers receive the same meal and the same delivery experience regardless of the platform choice. We select several restaurants that deliver in the area where the incentive aligned laboratory experiments are conducted and pair the restaurants so that one of the restaurants has diverging average ratings on the platforms, whereas the other has the same average rating on both platforms. In the sample of 223 restaurants that we collected in preparation for our experiments, 207 restaurants are rated on both Yelp and GrubHub and the average rating diverges by at least 0.5 points for 160 restaurants (77.3%) between the platforms. Among the restaurants chosen for our experiments, the restaurants with consistent ratings on both platforms serve as a control: it allows us to establish the subjects’ *a priori* preferences between Yelp and GrubHub brands. It also addresses a possible ’experimenter’s demand’ effect by obscuring from the subjects that we are interested in their choice between the platforms. We refer to these two types of restaurants as the Divergent Ratings Restaurant and the Control Restaurant.

To avoid branding or prior knowledge of a restaurant influencing the results, we do not display restaurant names to the subjects. We simply label the restaurant choices (rows) as Restaurant 1 and Restaurant 2 in the online data collection instrument that is implemented using Qualtrics platform. The example of the subjects’ decicion screen can be found in Appendix A. The real restaurant names are revealed only after the subjects complete the study including a demographic survey. For the purposes of reporting the results, we refer to the restaurants by nicknames that we assign based on their cuisine descriptions. In addition to the average restaurant rating, we provide the subjects with minimum information that would allow them to make an informed decision: the type of cuisine and the delivery hours (subjects are allowed to schedule a delivery for a later time during the day). For the stated preferences study, we keep the setting as similar as possible to the incentive aligned study; we show the subjects the same restaurant descriptions and delivery hours and ask them to imagine a scenario in which they want to order food from one of the two restaurants using either Yelp or GrubHub.

4.1 Treatments

We implement a between-subject design in which each subject faces only one combination of treatment variables and makes exactly one decision. The treatment variables are the ratings for two restaurants on two platforms (one rating for each platform), which are discussed below together with our randomization approach to deal with nuisance variables.

Focus Variable. The focus variable of our experimental design is the average restaurant rating. We choose the restaurants so that the Control Restaurant is rated 4.0 stars on both platforms, whereas the Divergent Ratings Restaurant has 4.5 stars on one platform and 3.5 stars on the other. We refer to the platform where the Divergent Ratings Restaurant has higher rating as the High Rating Platform and the platform where the restaurant has lower rating as the Low Rating Platform. To control for the possible interaction effect between the platform and the rating variables, we include two possible conditions: the High Rating Platform can be Yelp (*YelpHigh* treatment) or GrubHub (*GrubHubHig*h treatment). Note, again, that the quality of the food and the delivery will be exactly the same for the same restaurant regardless of platform choice. We emphasize this to the subjects on their decision screen, and if the subjects perceive this and act accordingly, their platform choice should be independent of the choice of the restaurant indicating no spillover effect from restaurant choice to platform choice.

In the incentive aligned study, we need to use information for actual restaurants that deliver to the location where the experiment is conducted. In *YelpHigh* condition, the Divergent Ratings Restaurant is Middle Eastern (ME) and the Control Restaurant is Sandwiches (SW) restaurant. In *GrubHubHigh* condition, the Divergent Ratings Restaurant is Pizza & Grill (PG) restaurant and the Control Restaurant is Coffee & Pizza (CP) restaurant. To summarize, in the incentive alignment study we have two treatments: *YelpHigh* and *GrubHubHigh*, which we implement by using two different combinations of local restaurants. In the stated preferences study, we are not tied to real restaurant information, but we aim to keep the setup as similar as possible to the incentive aligned study to make the results easily comparable. For this reason, we use the descriptions for Middle Eastern and Sandwiches restaurant pair from the incentive aligned study, but we randomly reassign the ratings in such a way that for some of the subjects the High Rating Platform is Yelp, and for the others it is GrubHub. Note that the random assignment of ratings to the restaurant–platform combination allows us to control for a possible interaction effect between the restaurant type and the ratings discrepancy, since the restaurants are equally likely to have diverging ratings in the stated preferences study.

Since the star rating is our focus variable that we manipulate between the platforms, we label the corresponding four treatments according to the restaurant–platform combination that has the highest 4.5 star rating: ME&G, ME&Y, SW&G, SW&Y in the stated preferences study. For example, in ME&G treatment, the Divergent Ratings restaurant is the Middle Eastern (ME) restaurant and the High Rating Platform is GrubHub (G). In other words, ME restaurant has 4.5 stars on GrubHub and 3.5 stars on Yelp, while the Sandwiches (SW) restaurant has 4.0 stars on both Yelp and GrubHub. We provide more details on the randomization in the next section, where we discuss how we handle nuisance variables.

Randomization. We are interested in whether the diverging average ratings affect platform choice even when they do not suggest any difference in the quality of the purchased item or the overall consumption experience that the consumer may expect to get. To study this, we ensure that the quality of the choice options stays constant regardless of the value of our focus variable, that is, the average rating associated with the restaurant on a specific platforms. However, there are a few other factors than the the average rating that can affect the choice of the restaurant–platform pair in our experimental design.

First, subjects can have *a priori* preference for one of the platforms or for a specific cuisine. We account for such preferences by observing the conditional probabilities of choosing one platform over another among those subjects who choose the Control Restaurant. Second, the order in which the restaurants and the platforms are presented in the layout of the data colletion instrument can influence the choices that the subjects make. Also, the restaurant category descriptions differ slightly between the two platforms, and Yelp stops accepting delivery orders 15 minutes earlier than GrubHub. To avoid picking up the confounding effects of such nuisance factors, we randomly shuffle them in the descriptions between the treatments as follows.

In the incentive aligned study, we can vary the order of the restaurants (either of the two restaurants is equally likely to be labelled as ’Restaurant 1’ or ’Restaurant 2’) and the order of the platforms (either of the two platforms is equally likely to be presented in the left or in the right column of the table) without compromising the realism of the treatments. In the stated preferences study, we can perform a comprehensive randomization by further varying the following factors. Between the platforms, we randomly swap the category descriptions for each of the two restaurants (four possible combinations), and the delivery hours (either Yelp or GrubHub closes 15 minutes earlier). We executed the randomization by implementing a real-time, random assignment of treatments to eliminate any systematic composition of subjects to a specific treatment, including the potential confounding effect of demographic factors.

4.2 Participants

The study subjects are recruited from three different pools. For the experiments using stated preferences method, we use (1) undergraduate business majors in a U.S. University who were offered a course credit for participating in the study during Fall 2019 semester, and (2) Amazon Mechanical Turk (mTurk) workers who received $2 reward for completing the survey in June 2020. These subjects participated in the stated preferences experiments online from their own private location, and we further set the requirement that mTurk workers must be located in the USA. For the incentive aligned study, we (3) recruited subjects using both announcements posted on the SONA system and printed advertising posted on the campus information boards. Any student, faculty or staff member were eligible to participate. The lab sessions for the incentive aligned study were conducted during Fall 2019 and Spring 2020 semesters, and the subjects received no other reward than the food order that they placed during the study.

5 Results

The finding show substantial support for the presence of a vertical spillover effect in platform choice. We first report the results from the stated preferences study and then from the incentive aligned experiment.

5.1 Stated Preferences Study

Since the stated preferences study involves two substantially different subjects pools with two different incentive types (course credit for undergraduate students and $2 reward for mTurk workers), we analyze and present the results for the two subject pools separately. In SONA pool (undergraduate students) we have 301 subjects, and in mTurk pool, we have 608 subjects. The subjects’ choices of our stated preferences study are summarized in Figure 1. First, we observe that despite the difference between the two subject pools, their choice patterns are remarkably similar, although the data from mTurk appears to be more noisy. We also observe that the subjects seem to prefer GrubHub over Yelp. This can be seen in all treatments as the subjects choose GrubHub as the Control Restaurant more often (in Figure 1, GrubHub is represented by yellow bars in left two columns, and by the green bars in right two columns). Second, we can visually confirm our hypothesis about the spillover effects. Recall that under the null hypothesis the subjects’ choice of the platform would be independent of the restaurant choice, and therefore the relative heights of the paired yellow and green bars in each panel should be the same. This is clearly not the case and we confirm the difference in proportions by chi-squared tests with p-values shown in each panel of Figure 1.

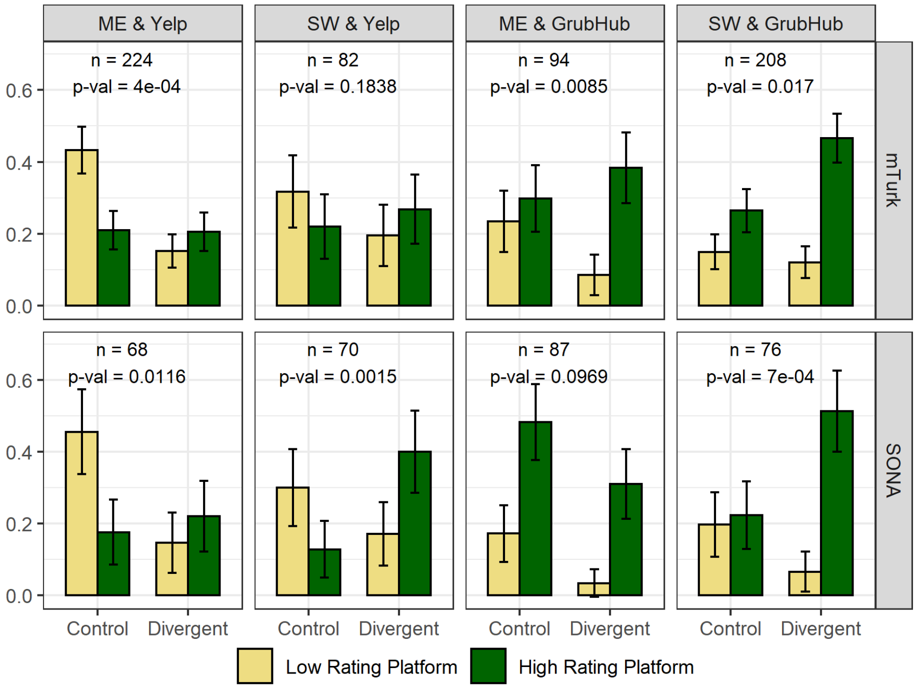


Figure 1. *Subject choices in stated preferences study*

5.2 Incentive Aligned Study

In the incentive aligned study, we have two treatments: *YelpHigh* and *GrubHubHigh* that have been described above. A total of 29 subjects participated in the study: 13 subjects were assigned to *YelpHigh* treatment and 16 subjects were assigned to *GrubHubHigh* treatment. Table 1 shows the number of subjects who chose each option, with the star rating for the corresponding platform–restaurant combination shown in brackets. We again observe that the subjects have *a priori* preference for GrubHub as many more subjects who choose the Control Restaurant (lower row) choose GrubHub. Those subjects who choose the Divergent Ratings Restaurant (upper row) again prefer to order from a platform where the restaurant is rated higher–particularly note how the subjects choose Yelp over Grubhub in *YelpHigh* treatment despite overwhelmingly preferring to use GrubHub when the restaurant has equal average ratings on both platforms (see lower row). The conditional probability estimates are, for *YelpHigh* treatment $\hat{p}_{Yelp|ME} = 0.714$, which is greater than $\hat{p}_{Yelp|SW} = 0$, and for *GrubHubHigh* treatment, $\hat{p}_{Yelp|PG} = 0.091$, which is smaller than $\hat{p}_{Yelp|CP} = 0.2$. The results are consistent with the presence of vertical spillover effect.

	YelpHigh			*GrubHubHigh*	
	Yelp	GrubHub		Yelp	GrubHub
Middle Eastern	5 (4.5 stars)	2 (3.5 stars)	Pizza & Grill	1 (3.5 stars)	10 (4.5 stars)
Sandwiches	0 (4 stars)	6 (4 stars)	Coffee & Pizza	1 (4 stars)	4 (4 stars)

Table 1. *Platrform choice in the incentive aligned study.*

To test for the independence between the restaurant choice and the platform choice statistically, we use Pearson’s Chi-squared test with p-values obtained by Monte Carlo simulation with 106 replicates. The result is significant for *YelpHigh* treatment (p-value = 0.0210), but not significant for *GrubHubHigh* (p-value ≈ 1). However, note that in the *GrubHubHigh* treatment, it is statistically much harder to detect a significant difference due to a strong preference for GrubHub platform among the subjects; we have only two subjects who chose Yelp platform in the treatment. However, since the results of *GrubHubHigh* treatment do not contradict the results we observed in *YelpHigh* treatment, and they are consistent with what we observe in the stated preferences studies, we conclude that the incentive aligned study supports the presence of a vertical spillover effect.

5.3 The Effect of the Number of Ratings

In our main experiment we wanted to isolate the effect of average rating, and hence, in the experiment we exclude as many nuisance factors as possible, including the number of ratings on which the average is based. However, when choosing the restaurants for our study we observed that the rating discrepancy between platforms often occurs when the restaurant is new to the platform, and therefore has not yet had time to accumulate many ratings. In this case, it is reasonable to believe that the diverging average ratings across platforms is explained by variation associated with a small sample size. If the subjects perceive that the small sample is small, the number of ratings should mitigate the spillover effect that we observe in our experiments. Therefore, we conduct an additional stated preferences study in which we try to establish if the number of ratings affects our subject's choices. We present the subjects with the following scenario:

> *Imagine that you are offered a choice between $3 in cash and a [g] gift card towards delivery from a certain restaurant ("Restaurant X"). On Yelp, Restaurant X has an average rating of [r] stars, based on [n] ratings What do you choose?*

Here, our focus variable is the number of ratings n, which can take three possible values: 1, 3 or 15. It seems likely that the direction of the effect of the number of ratings is different for high and low rated restaurant, and to control for this, we have Low Rating Condition where the average rating $r = 2.0$, and High Rating Condition where the average rating $r = 4.0$. To ensure that sufficient percentage of subjects select each option, we adjust the corresponding values of the gift card to be $g = \$50$ in Low Rating Condition and $g = \$20$ in High Rating Condition. The values of the gift cards and cash were established during a pilot study in a way that would cause a reasonable proportion breakdown between the subjects who choose cash and the subjects who choose a gift card.

For the average rating condition, we implement a within-subject design, where part of the subject faces the Low Rating Condition first, and another part faces the High Rating Condition first. In each of these two conditions the number of ratings n is drawn independently from one of the three values (1, 3 or 15). To summarize, we have a full factorial design with 2 (Average rating condition) x 3 (Number of ratings condition) and a total of six treatments.

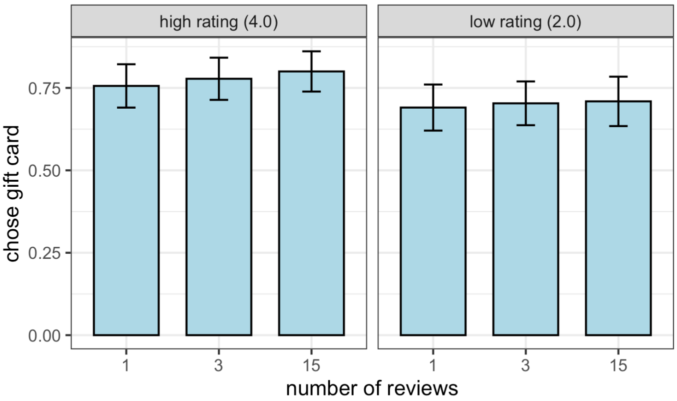


Figure 2. The impact of the number of ratings on the choice between a gift card and a cash reward.

In this study, we have a total of 491 subjects recruited through mTurk, each participating in both Average Rating conditions in a random order. Figure 2 shows the percentage of subjects who chose the restaurant gift card in each of the treatments. In both average rating conditions the percentage of subjects who choose the restaurant slightly increases with the number of reviews, but the differences are statistically insignificant. Furthermore, in Low Rating condition the direction of change goes against our expectations as more subjects would seem to be confident to choose a restaurant with a low rating as the nuber of ratings increases and thus confirms the accuracy of the low average rating. However, if differences exist, they are very small: even if we compare the treatments with 1 and 15 ratings, the 95% confidence interval between the proportion of subjects who chose the gift card is between and -0.084 and 0.121 in Low Average Rating condition and between -0.046 and 0.136 for High Average Rating condition. We conclude that in the food delivery platform setting the number of ratings appears to have little to no effect on the probability of the subjects choosing the restaurant rather than opting out and taking the cash. This suggests that diverging average ratings across platforms can have an impact on platform choice even if they result from variance in small samples.

6 Discussion and Conclusions

Average ratings have a substantial impact on consumer behavior on platforms and they are believed to affect competition between platforms (e.g. Chevalier and Mayzlin 2006, Kathuria and Lai 2018). Yet, a few studies have identified mechanisms by which the familiar 5-star ratings system shapes competition between platforms and, to this end, our study focuses the impact of diverging average ratings across platforms on platform choice. Our experiments show that consumers are more likely to buy an item from a platform where it is rated higher on average, even if the ratings do not inform about the relative merits of competing platforms. We explain this by a vertical spillover effect that takes place along the value chain as the evaluations of vendors (or their items) spill over and affect consumers' perceptions of the platform itself. The results are intuitive and allow us to make contributions to literature on spillover effects, platform competition, and online ratings. We also reflect upon methodological differences between the incentive aligned and stated preferences studies in the context of online ratings, and discuss managerial and regulatory implications of our results.

6.1 Contributions

Previous studies have argued that better valence and volume of average ratings can provide a competitive advantage to a retail platform (Chevalier and Mayzlin 2006, Kathuria and Lai 2018), whereas others have observed that ratings tend to become inflated over time (Athey et al. 2019, Filippas et al. 2018, Hu et al. 2009, Kokkodis 2021, Nosko and Tadelis 2015, Zervas et al. 2020), but no study has shown how the two phenomena are empirically connected.

Vertical Spillover Effects. Spillover effects have been studied as a horizontal phenomenon in which influence spills over to brands, companies, or product categories that occupy the same position as the focal entity in the value chain. By contrast, we have shown that spillover effects can also take place vertically between entities that occupy different positions in the value chain, and how such a vertical spillover effect can result in broader implications such as driving ratings inflation when both platform vendors and consumers multihome. It is important to note that consumers do not need to believe that they get a better item or the overall consumption experience by buying from a platform where the item is rated higher. They just need to assume that they cannot be worse off by buying from a platform where the item is rated higher. For instance, the study by Li and Agarwal (2017) suggest that there are other types of vertical spillover effects taking place in digital platforms that are not well understood at the moment.

Platform Competition. A platform that has higher average ratings for the same items than its competitor can 'steal' transactions from the competitor that provides perhaps more useful information and, hence, a sort of virtual showrooming to consumers. Yet, consumers may also feel safer to choose to eventually buy from a platform where an item is rated higher due to the halo of diverging average ratings on the platforms. Again, it is important to stress that our experimental design controls for i) the necessary platform attributes (food category, restaurant delivery hours) that are part of the treatment, and for ii) *a priori* preference for GrubHub over Yelp among our subjects. The results mean that providing more accurate but lower average ratings can be disadvantageous under platform competition. This may consequently incentivize the platform owner to let the ratings become inflated. The findings align with recent literature on the impact of fake reviews in that inflated, that is, less accurate ratings may be beneficial to a platform in the short term (Wang et al. 2020). However, in the long term ratings inflation will erode the usefulness of the standard 5-star system and, perhaps, render it eventually obsolete as recently speculated by Filippas et al. (2018, p. 1): *”as the potential to harm is what makes ratings effective, reputation systems, as currently designed, sow the seeds of their own irrelevance.”*

Online Ratings. Previous studies have identified vendor manipulation as one of the main reasons for ratings inflation (Fradkin et al. 2015, Hu et al. 2009, Lee and Kai 2020), but they have not explained why platform owners are sometimes interested in letting the inflation happen. We show that a platform can gain competitive advantage from higher average ratings of its vendors and their items due to a vertical spillover effect, which motivates the platform owner to allow vendors to take actions that can drive ratings inflation. Ironically, if combating ratings manipulation is costly as one would expect, then a platform that aims to offer more accurate and not inflated ratings may end up investing in operations that hurt itself in competition with those who adopt a more relaxed policy toward vendors who manipulate their ratings.

Finally, it is somewhat surprising that the number of ratings does not seem to mitigate the impact of diverging average ratings across the platforms in our study, adding to the mixed results about the impact of ratings volume on consumer behavior (Blal and Sturman 2014, Watson et al. 2018, Zimmermann et al. 2018). Our empirical results do not allow to say more about the impact of the number of ratings than it would seem likely that the impact varies substantially from a setting to another and that the issue may require further study.

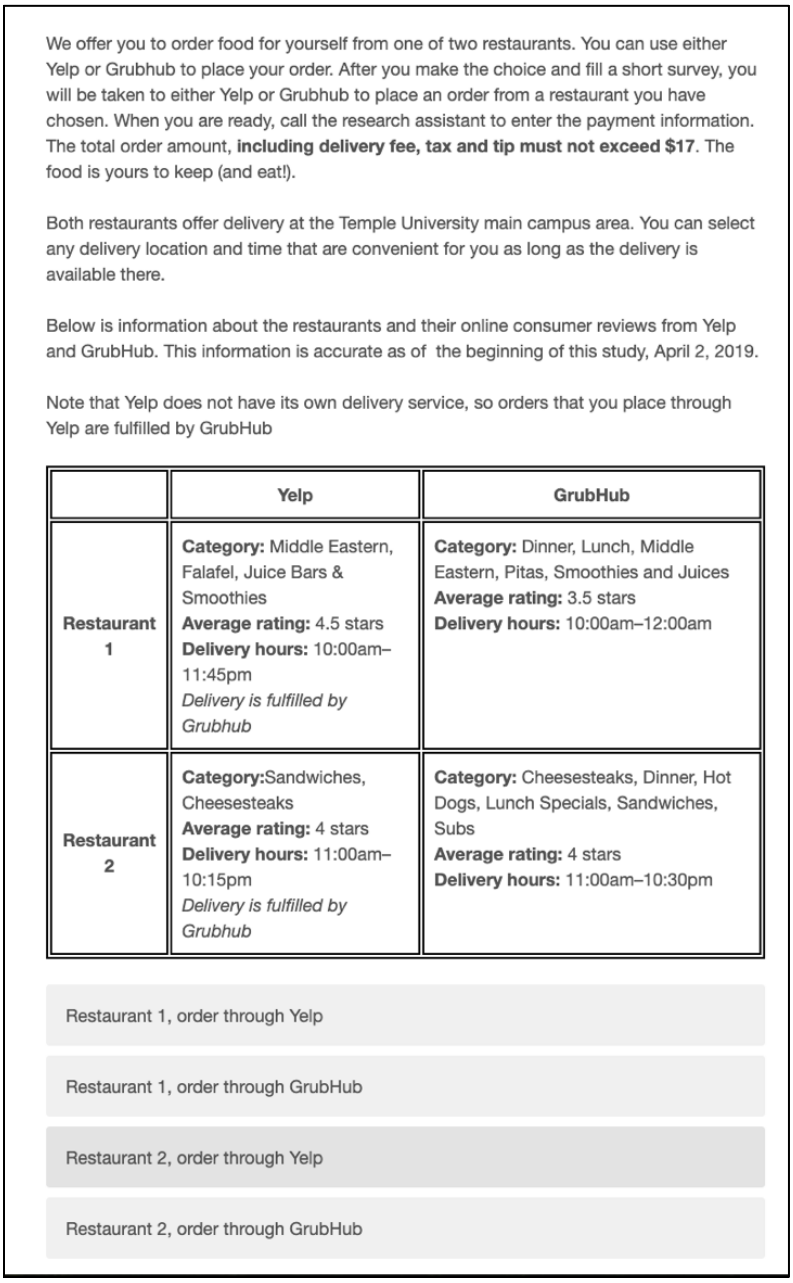
6.2 Managerial and Regulatory Implications

Increasing competition between platforms that use the standard 5-star ratings system can drive platform owners to govern the system in a way that results in ratings inflation. This should alert both consumers and regulators. The former need to become increasingly critical of information provided by the 5-star system and average ratings in particular as they may have become so inflated that the ratings do not anymore usefully distinguish the rated items. This is what has largely happened, for instance, on Uber platform whose drivers typically have an average rating close to the maximum. More generally, while competition generally benefits consumers, platform owners may have little incentive to fight ratings inflation under increasing competition between platforms. In this sense, our conjecture parallels the findings of Wang et al. (2020) who show analytically that positive fake reviews benefit platform in the short term, while the long-term implications are difficult to predict. Yet, it is not clear what collective mechanisms may be put in place to curb ratings inflation and to safeguard the integrity of the standard 5-star ratings system. If the system becomes obsolete due to ratings inflation (Filippas et al. 2018), search costs for consumers will increase. This may require platforms to innovate new kinds of ratings systems that stand out from the competition and help reduce consumers’ search costs, whereas the suggested portability of ratings and reviews (Kathuria and Lai 2018) makes sense only if platforms use a standardized system for collecting and displaying product evaluations. The risk is currently that ratings inflation will start to negatively impact consumers’ capacity to make successful purchase decisions and to erode consumer trust in the 5-star system.

6.3 Methodological Reflections

The way our study combines the stated preferences method and the incentive aligned experiments allows us to also discuss the relative merits of the two methods for studying consumer reactions to online ratings. The stated preferences method is widely used in marketing and information systems research as well as in the industry, where ’would you’ type customer surveys often provide input for important marketing and design decisions. The method is cheap and can be used where experiments involving the observation of real behavior are not feasible. However, there is a concern that the preferences expressed by survey respondents differ from their real-world behavior and, consequently, it is recommend that one should not rely on stated preferences data unless it has been shown that behavior in the setting can be inferred using hypothetical incentives (Katok 2011). For instance, subjects in a stated preferences study may try maximize their utility by completing the experimental task as quickly as possible, without taking time to carefully consider the information provided. A study based on stated preferences could therefore mistakenly conclude that consumers fail to realize that the difference in the average rating for the same product on different platforms should be attributed to randomness, whereas they were simply not motivated to consider the problem setup carefully. Our findings are consistent across the two methods, which offers an additional robustness check to our study and also suggests that the stated preferences method is a valid approach in our context.

Appendix A: Subject's Screen in the Incentive Aligned Study



We offer you to order food for yourself from one of two restaurants. You can use either Yelp or Grubhub to place your order. After you make the choice and fill a short survey, you will be taken to either Yelp or Grubhub to place an order from a restaurant you have chosen. When you are ready, call the research assistant to enter the payment information. The total order amount, **including delivery fee, tax and tip must not exceed $17**. The food is yours to keep (and eat!).

Both restaurants offer delivery at the Temple University main campus area. You can select any delivery location and time that are convenient for you as long as the delivery is available there.

Below is information about the restaurants and their online consumer reviews from Yelp and GrubHub. This information is accurate as of the beginning of this study, April 2, 2019.

Note that Yelp does not have its own delivery service, so orders that you place through Yelp are fulfilled by GrubHub

	Yelp	**GrubHub**
Restaurant 1	**Category:** Middle Eastern, Falafel, Juice Bars & Smoothies **Average rating:** 4.5 stars **Delivery hours:** 10:00am–11:45pm *Delivery is fulfilled by Grubhub*	**Category:** Dinner, Lunch, Middle Eastern, Pitas, Smoothies and Juices **Average rating:** 3.5 stars **Delivery hours:** 10:00am–12:00am
Restaurant 2	**Category:**Sandwiches, Cheesesteaks **Average rating:** 4 stars **Delivery hours:** 11:00am–10:15pm *Delivery is fulfilled by Grubhub*	**Category:** Cheesesteaks, Dinner, Hot Dogs, Lunch Specials, Sandwiches, Subs **Average rating:** 4 stars **Delivery hours:** 11:00am–10:30pm

Restaurant 1, order through Yelp
Restaurant 1, order through GrubHub
Restaurant 2, order through Yelp
Restaurant 2, order through GrubHub

References

Ananthakrishnan UM, Li B, Smith MD (2020) A tangled web: should online review portals display fraudulent reviews? Information Systems Research 31(3):950–971.

Athey S, Castillo JC, Chandar B (2019) Service quality in the gig economy: empirical evidence about driving quality at Uber. SSRN.

Aziz A, Li H, Telang R (2020) The consequences of rating inflation on platforms: evidence from a quali-experiment. SSRN URL https://ssrn.com/abstract=3842174.

Blal I, Sturman MC (2014) The differential effects of the quality and quantity of online reviews on hotel room sales. Cornell Hospitality Quarterly 55(4):365–375.

Borah A, Tellis GJ (2016) Halo (spillover) effects in social media: Do product recalls of one brand hurt or help rival brands? Journal of Marketing Research 53(2):143–160.

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

Chevalier JA, Dover Y, Mayzlin D (2018) Channels of impact: User reviews when quality is dynamic and managers respond. Marketing Science 37(5):688–709.

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

de Langhe B, Fernbach PM, Lichtenstein DR (2016) Navigating by the stars: investigating the actual and perceived validity of online user ratings. Journal of Consumer Research 42:817–833.

Filippas A, Horton JJ, Golden J (2018) Reputation Inflation (Ithaca, NY, USA).

Financial Times (2018) Apple: how app developers manipulate your mood to boost ranking. https: //www.ft.com/content/217290b2-6ae5-47f5-b1ac89c6ccebab41

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Information Systems Research 19(3):291–313.

Fradkin A, Grewal E, Holtz D, Pearson M (2015) Bias and reciprocity in online reviews: evidence from field experiments on Airbnb. Proceedings of the Sixteenth ACM Conference on Economics and Computation.

Gao G, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minorty and silent majority: How do do online ratings reflect population perceptions of quality. MIS Quarterly 39(3):565–589.

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

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

Hu N, Pavlou PA, Zhang J (2009) Overcoming the J-shaped distribution of product reviews. Communications of the ACM 52(10):144–147.

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

Huotari P, Jarvi K, Kortelainen S, Huhtamaki J (2017) Winner does not take all: selective attention and local bias in platform-based markets. Technological Forecasting & Social Change 114:313–326.

Ibbotson, A (2018) Patients trust online reviews as much as doctor recommendations—and other shocking facts about transparency in healthcare. https://nrchealth.com/patients-trust-online-reviews/

Jabr W, Zheng ZE (2014) Know yourself and know your enemy: an analysis of firm recommendations and consumer reviews in a competitive environment. MIS Quarterly 38(3):635–654.

Janakiraman R, Sismeiro C, Dutta, S (2009) Perception spillovers across competing brands: A disaggregate model of how and when. Journal of Marketing Research 46(4):467--481.

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

Kathuria V, Lai JC (2018) User review portability: why and how? Computer Law & Security Review 34:1291–1299.

Katok E (2011) Using laboratory experiments to build better operations management models. Foundations and Trends in Technology, Information and Operations Management, 5(1):1–86.

Katsamakas E, Madany H (2019) Effects of user cognitive biases on platform competition. Journal of Decision Systems 28(2):138–161.

Kim B, Lee Jark H (2017) Two-sided platform competition with multihoming agents: An empirical study on the daily deals market. Information Economics and Policy 41(Dec):36–53.

Kokkodis M (2021) Dynamic, multidimensional, and skillset-specific reputation systems for online work. Information Systems Research Articles in advance.

Krijestorac H, Garg R, Mahajan V (2020) Cross-platform spillover effects in consumption of viral content: a quasi-experimental analysis using synthetic controls. Information Systems Research 31(2):449–472.

Kumar N, Qiu L, Kumar S (2018a) Exit, voice, and response on digital platforms: an empirical investigation of online management response strategies. Information Systems Research 29(4):849–870.

Kumar N, Venugopal D, Qiu L, Kumar S (2018b) Detecting review manipulation on online platforms with hierarchical supervised learning. Journal of Management Information Systems 35(1):350–380.

Kumar N, Venugopal D, Qiu L, Kumar S (2019) Detecting anomalous online reviewers: an unsupervised approach using mixture models. Journal of Management Information Systems 36(4):1313–1346.

Lee G, Kai J (2020) Comparing numerical ratings and plain-text feedback from online reputation system: evidence from sentiment analysis of Airbnb reviews in London. SSRN URL https://ssrn.com/ abstract=3611064.

Lee MK, Kusbit D, Metsky E, Dabbish L (2015a) Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea).

Lee YJ, Hosanagar K, Tan Y (2015b) Do I follow my friends or the crowd? Information cascades in online movie ratings. Management Science 61(9).

Li X (2018) Impact of average rating on social media endorsement: the moderating role of rating dispersion and discount threshold. Information Systems Research 29(3):739–754.

Li Z, Agarwal A (2017) Platform integration and demand spillovers in complementary markets: evidence from Facebook's integration of Instagram. Management Science 63(10):3438–3458.

Moe WW, Schweidel DA (2012) Online product opinions: incidence, evaluation, and evolution. Marketing Science 31(3).

Nosko C, Tadelis S (2015) The limits of reputation in platform markets: an empirical analysis and field experiment. Technical Report 20830, National Bureau of Economic Research.

Pagano D, Maalej W (2013) User feedback in the AppStore: an empirical study. 125–134 (Rio de Janeiro, Brasil).

Rietveld J, Schilling MA (2020) Platform competition: a systematic and interdisciplinary review of the literature.

Rocklage MD, Fazio RH (2020) The enhancing versus backfiring effects of positive emotion in consumer reviews. Journal of Marketing Research 57(2):332–352.

Roehm ML, Tybout AM (2006) When will a brand scandal spill over, and how should competitors respond? Jounral of Marketing Research 43(3):366–373.

Sahni NS (2016) Advertising spillovers: evidence from online field experiments and implications for returns on advertising. Journal of Marketing Research 53:459–478.

Sahoo N, Dellarocas C, Srinivasan S (2018) The impact of online product reviews on product returns. Information Systems Research 29(3):723–738.

Schilling MA (2002) Technology success and failure in winner-take-all markets: the impact of learning orientation, timing, and network externalities. Academy of Management Journal 45(2):387–398.

Shen W, Hu YJ, Ulmer JR (2015) Competing for attention: an empirical study of online reviewers' strategic behavior. MIS Quarterly 39(3):683–696.

Spiegel Research Center (2017). How online reviews influence sales. https://spiegel.medill.northwestern.edu/online-reviews

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

Tunc MM, Cavusoglu H, Raghunathan S (????) Online product reviews: is a finer-grained rating scheme superior to a coarser one? MIS Quarterly.

Wang Z, Kumar S, Liu D (2020) On platform's incentive to filter fake reviews: a game-theoretic model. Proceedings 41st International Conference on Information Systems (ICIS) (Hyderabad, India).

Watson J, Ghosh AP, Trusov M (2018) Swayed by the numbers: the consequences of displaying product review attributes. Journal of Marketing 82(6):109–131.

Womply (2018). How online reviews impact small business revenue. https://www.womply.com/impact-of-online-reviews-on-small-business-revenue/

Xu AJ, Schwarz N (2018) How one thing leads to another: spillover effects of behavioral mind-sets. Current Directions in Psychological Science 27(1):51–55.

Yin D, Mitra S, Zhang H (2016) When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. Information Systems Research 27(1):131–144.

Zervas G, Proserpio D, Byers JW (2020) A first look at online reputation on Airbnb, where every stay is above average. SSRN URL https://ssrn.com/abstract=2554500.

Zimmermann S, Herrmann P, Kundisch D, Nault BR (2018) Decomposing the variance of consumer ratings and the impact on price and demand. Information Systems Research 29(4):984–1002.